Mathematics of Deep Learning
Stat 946 Fall 2024

Aukosh Jagannath

Department of Statistics and Actuarial Science, University of Waterloo
Email address: a.jagannath@uwaterloo.ca

This research is funded by NSERC.

Abstract. These are the course notes for Stat 946 in Fall 2024.

Contents

Chapter 1. tba 1

Chapter 2. tba 2

Chapter 3. tba 3

Chapter 4. Implicit Regularization 4
1. Introduction 4
2. A Classification Example 5

Chapter 5. Interpolation does not imply poor generalization 9
1. Introduction 9
2. Non-Parametric Regression (Kernel Smoothing) 10
3. Double Descent 13
4. Appendix 14

Chapter 6. A Primer on Random Matrix Theory 15
1. Introduction 15
2. Marchenko-Pastur Law 16
3. Proof of the Marchenko–Pastur Law 18
4. Appendix 21

Chapter 7. Double Descent 23
1. Ridgeless Least Squares 23
2. Double Descent 25
3. Appendix 28

Chapter 8. Neural Tangent Kernel 29
1. Overview: Training Dynamics 29
2. Lazy Training 30
3. Kernel Methods 31
4. Neural Tangent Kernel 32

Chapter 9. Lazy Training 36
1. Introduction 36
2. Appendix: Theorems and Definitions 39

Chapter 10. TBA 40

Chapter 11. Sample Complexity of SGD 41
1. Introduction 41

2. Applications to Some Inference Problems 42
3. Preliminaries 44
4. Information Exponent 46
5. Proof Idea 47

Chapter 12. Effective Dynamics 49
1. Introduction 49
2. Preliminaries 50
3. Effective Dynamics 51
4. Application to Some High-Dimensional Problems 53

Chapter 13. Spectral Alignment 56
1. Introduction 56
2. Summary of technicalities and proofs 64

Chapter 14. Transformers 66
1. Introduction 66
2. The master equation and continuity-based analysis 69

Bibliography 73

CHAPTER 1

tba

CHAPTER 2

tba

CHAPTER 3

tba

CHAPTER 4

Implicit Regularization

1. Introduction

What happens when you overparameterize a model? What happens if the hypothesis class $\mathcal{F}$ contains an interpolator, i.e. there exists some $\hat{f} \in \mathcal{F}$ such that $\hat{f}$ interpolates (perfectly fits the training data), i.e. $\hat{R}(\hat{f}) = 0$. In classical learning theory, achieving zero training error ($\hat{R}(\hat{f}) = 0$) usually implies that for sure that the generalization error will be bad $R(\hat{f}) \neq 0$. However, in modern deep learning, overparameterization can actually *improve* generalization in practice.

The observation from various empirical studies is that overparameterized models often generalize well. In these 2 lectures, we will explore several research works that attempt to explain this phenomenon. A key concept that may explain this is **implicit regularization**: the optimization algorithm used during training has a bias towards special solutions with desirable properties, even without explicit regularization.

Example 4.1 (Overparameterized Least Squares). Given $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$, consider the problem of minimizing the squared loss:

$$\min_{\theta \in \Theta \subseteq \mathbb{R}^d} \underbrace{\|X\Theta - y\|_2^2}_{\hat{R}_n}$$

where our assumption here is that $d > n$ (overparameterized). The solution space (the minimizer) is some affine subset of $\mathbb{R}^d$ with codimension n. We use gradient descent with step size η_t to solve this optimization problem, which is an iterative scheme of the form

$$\begin{cases} \theta_{t+1} & = \theta_t - \eta_t \nabla \hat{R}_n(\theta_t) \\ \theta_0 & = 0 \end{cases}$$

where we have $\hat{R}_n(\theta) = \frac{1}{n}\|X\theta - y\|_2^2$. Which solution does this algorithm pick?

Lemma 4.2. *$\theta_t \to \hat{\theta}$, which solves*

$$\min_{\theta \in \Theta} \|\theta\|^2 \text{ subject to } X\theta = y,$$

i.e. the minimum norm interpolant.

Proof. The KKT (Karush-Kuhn-Tucker) conditions for this optimization problem are

$$X\theta = y \quad \text{and} \quad 2\theta + X^\top \mu = 0$$

where μ is the lagrange mutliplier. The KKT conditions will hold for any algorithms that:

(1) interpolates the data

(2) stays within the row space of X, $\text{Row}(X)$.

□

2. A Classification Example

Next, we consider the classification problem with the focus on linearly separable data. Is this a reasonable assumption? For today's lecture, we will work under this assumption.

DEFINITION 4.3. A set $E \subseteq \mathbb{R}^d \times \{\pm 1\}$ is **linearly separable** if $\exists w \in \mathbb{R}^d$ such that $y\langle w, x\rangle > 0$ for all $(y, x) \in E$.

Assumption. $|E|$ is finite, with $E = \{(y_1, x_1), \ldots, (y_n, x_n)\}, |E| = n$ and is linearly separable.

DEFINITION 4.4. If E is linearly separable, the L^2 **max margin predictor** is the unique solution to

$$\min_w \ \|w\|_2 \quad \text{subject to } y_i\langle w, x_i\rangle \geq 1, \forall i.$$

Let $\overline{w}$ be the L^2 max margin predictor. The **margin** is defined as $\gamma = \min_j y_j\langle \overline{w}, x_j\rangle$ and the **support vectors** are defined as $\mathcal{S} = \arg\min_i \langle \overline{w}, x_i\rangle$.

REMARK 4.5. The dual of the optimization problem is $\max_{\|u\|_2=1} \min y_i\langle u, x_i\rangle$.

Here is the setup we will follow: we are given the dataset $\{(y_i, x_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ are the feature vectors, and $y_i \in \{\pm 1\}$ are the binary labels. The loss function is defined as:

$$\mathcal{L}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i\langle w, x_i\rangle),$$

where $\ell(\cdot)$ is the chosen loss function.

EXAMPLE 4.6. Here are some commonly used loss functions:

- Logistic Loss: $\ell(t) = \log(1 + e^{-t})$
- Exponential Loss: $\ell(t) = e^{-t}$

THEOREM 4.7 (Soudry et al. [2018], simplified). *For almost all linearly separable datasets, if the step size satisfies*

$$\eta < \frac{1}{c \cdot \lambda_{\max}^2\left(\frac{1}{n}xx^\top\right)},$$

for some constant c and anu starting point $w(0)$, the gradient descent iterates w_t will behave as

$$w_t = \hat{w}\log(t) + \rho(t),$$

where $\hat{w}$ is the L^2 max margin predictor, and $\|\rho(t)\| = O(\log\log t)$.

REMARK 4.8. In classification, the focus is not on the specific value of $w(t)$, but rather on the direction of $w(t)$, i.e., $\frac{w(t)}{\|w(t)\|}$. In particular, this theorem gives us

$$\frac{w(t)}{\|w(t)\|} \sim \frac{\hat{w}}{\|\hat{w}\|}\left(1 + o(1)\right), \text{and so } \lim_{t\to\infty}\frac{w(t)}{\|w(t)\|} = \frac{\hat{w}}{\|\hat{w}\|},$$

meaning that the direction of $w(t)$ asymptotically aligns with the direction of the max-margin predictor $\hat{w}$.

PROOF. Consider the gradient flow, which is the continuous-time limit of gradient descent, i.e., taking the step size $\eta \to 0$ appropriately. The dynamics are given by:

$$\begin{cases} \dot{w}(t) = -\nabla \mathcal{L}(w(t)) \\ w(0) = 0 \end{cases}$$

The goal is to ensure that $\|w(t) - \hat{w}\log(t)\|$ remains small. To achieve this, define: $r(t) = w(t) - \hat{w}\log(t) - \widetilde{w}$, and the objective is to bound $\|r(t)\| \leq C$ for some constant C so that it's small. Here, $\widetilde{w}$ is chosen such that:

$$\sum_{i \in \mathcal{S}(\hat{w})} \exp(-\langle \widetilde{w}, x_i \rangle) x_i = \hat{w}.$$

Why does this $\widetilde{w}$ exist?
Recall the original optimization problem:

$$\min_w \|w\| \quad \text{subject to} \quad y_i \langle w, x_i \rangle \geq 1 \text{ for all } i.$$

This is a convex optimization problem. By applying the Lagrange multiplier method, we can derive the optimality conditions.
Using the KKT conditions, we form the Lagrangian:

$$\mathcal{H}(w, \mu) = \|w\|^2 + \sum_i \mu_i \left(y_i \langle w, x_i \rangle - 1 \right),$$

where $\mu_i \geq 0$ are the dual variables. The optimality conditions are:

$$\begin{aligned} \mu_i &\geq 0, \forall i \\ y_i \langle w, x_i \rangle &\geq 1, \forall i \\ \hat{w} &= \sum_{i \in \mathcal{S}} a_i x_i \qquad \text{(Stationarity Condition, i.e. } \nabla \mathcal{H} = 0\text{)}. \end{aligned}$$

Pick $\tilde{w}$ such that $\exp(-\langle \tilde{w}, x_i \rangle) = a_i$, i.e. $\tilde{w} X = -\log a$, and we are done.
Fact. For almost every dataset, $|\mathcal{S}| \leq d$.

We would now like to prove that $\|r_t\| \leq C$ is small. We want to control $\|r_t\|$. Now,

$$\dot{r} = \dot{w} - \frac{1}{t}\hat{w} = -\nabla \mathcal{L} - \frac{1}{t}\hat{w}$$

We have

$$\begin{aligned} -\nabla \mathcal{L} &= -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial w} \left[e^{-y_i \langle w, x_i \rangle} \right] \\ &= \frac{1}{n} \sum_{i=1}^{n} y_i x_i e^{-y_i \langle w, x_i \rangle} \end{aligned}$$

and thus by letting $\overline{x}_i = y_i x_i$, we have $\dot{r} = \frac{1}{n}\sum_{i=1}^n \overline{x}_i e^{-\langle \overline{x}_i, w\rangle} - \frac{1}{t}\hat{w}$. Therefore,

$$\begin{aligned}\dot{\left(\frac{1}{2}\|r\|^2\right)} &= \langle r, \dot{r}\rangle \\ &= \left\langle r, \sum_{i=1}^n \exp(-\langle w, \overline{x}_i\rangle)\overline{x}_i - \frac{1}{t}\hat{w}\right\rangle \\ &\overset{(*)}{=} \left\langle r, \sum_i \exp\left(-\log(t)\langle \hat{w}, \overline{x}_i\rangle - \langle \widetilde{w}, \overline{x}_i\rangle - \langle r, \overline{x}_i\rangle\right) x_i - \frac{1}{t}\hat{w}\right\rangle\end{aligned}$$

where for $(*)$, we use the fact that $r(t) = w(t) - \log(t)\hat{w} - \widetilde{w}$.
Finally, we will decompose the last sum over support vectors $\mathcal{S}$ and non-support vector

$$\begin{aligned}\text{(I)} &= \left\langle r, \sum_{i\in\mathcal{S}} \exp\left(-\log(t)\langle \hat{w}, \overline{x}_i\rangle - \langle \widetilde{w}, \overline{x}_i\rangle - \langle r, \overline{x}_i\rangle\right) x_i - \frac{1}{t}\hat{w}\right\rangle \\ \text{(II)} &= \left\langle r, \sum_{i\in\mathcal{S}^c} \exp\left(-\log(t)\langle \hat{w}, \overline{x}_i\rangle - \langle \widetilde{w}, \overline{x}_i\rangle - \langle r, \overline{x}_i\rangle\right) x_i - \frac{1}{t}\hat{w}\right\rangle\end{aligned}$$

A quick observation to note here is that on $\mathcal{S}$, we have $\langle \hat{w}, \overline{x}_i\rangle = 1$ and recall our definition of $\widetilde{w}$, we have $\hat{w} = \sum_{i\in\mathcal{S}} \exp(-\langle \widetilde{w}, x_i\rangle)x_i = \sum_{i\in\mathcal{S}} \exp(-\langle \widetilde{w}, \overline{x}_i\rangle)\overline{x}_i$.

Therefore, we may rewrite our expression in (I) as:

$$\begin{aligned}\text{(I)} &= \frac{1}{t}\left[\sum_{i\in\mathcal{S}} \exp(-\langle \widetilde{w}, \overline{x}_i\rangle - \langle r, \overline{x}_i\rangle)\langle r, \overline{x}_i\rangle - \sum_{i\in\mathcal{S}} \exp(-\langle \widetilde{w}, \overline{x}_i\rangle)\rangle\langle r, \overline{x}_i\rangle\right] \\ &= \frac{1}{t}\sum_{i\in\mathcal{S}} \exp(-\langle \widetilde{w}, \overline{x}_i\rangle) \cdot \langle r, \overline{x}_i\rangle \cdot (\exp(-\langle r, \overline{x}_i\rangle) - 1) \\ &\leq 0\end{aligned}$$

since $z(e^{-z} - 1) \leq 0$ for all $z \in \mathbb{R}$.

Now, for the non-support vectors: note that $e^{-z}z \leq 1$, and $\psi = \min_{i\in\mathcal{S}^c}\langle \overline{x}_i, \hat{w}\rangle > 1$. Therefore,

$$\begin{aligned}\text{(II)} &\leq \sum_{i\in\mathcal{S}^c} \exp[-\log(t)\langle \hat{w}, \overline{x}_i\rangle - \langle \widetilde{w}, \overline{x}_i\rangle] \cdot \exp(-\langle \overline{x}_i, r\rangle) \cdot \langle \overline{x}_i, r\rangle \\ &\leq \frac{1}{t^\psi}\sum_{i\in\mathcal{S}^c} \exp(-\langle \widetilde{w}, \overline{x}_i\rangle)\end{aligned}$$

where $C := \sum_{i\in\mathcal{S}^c} \exp(-\langle \widetilde{w}, \overline{x}_i\rangle)$ is a constant independent of t. Thus, we have

$$\dot{\left(\frac{1}{2}\|r\|^2\right)} \lesssim \frac{C}{t^\psi}$$

Since $\psi > 1$, we have that for all $t > s$,

$$\|r(t)\|^2 - \|r(s)\|^2 \leq C\int_s^t \frac{dt}{t^\psi} < \infty.$$

□

CHAPTER 5

Interpolation does not imply poor generalization

1. Introduction

There is a folklore in statistical learning theory: "interpolation is bad."

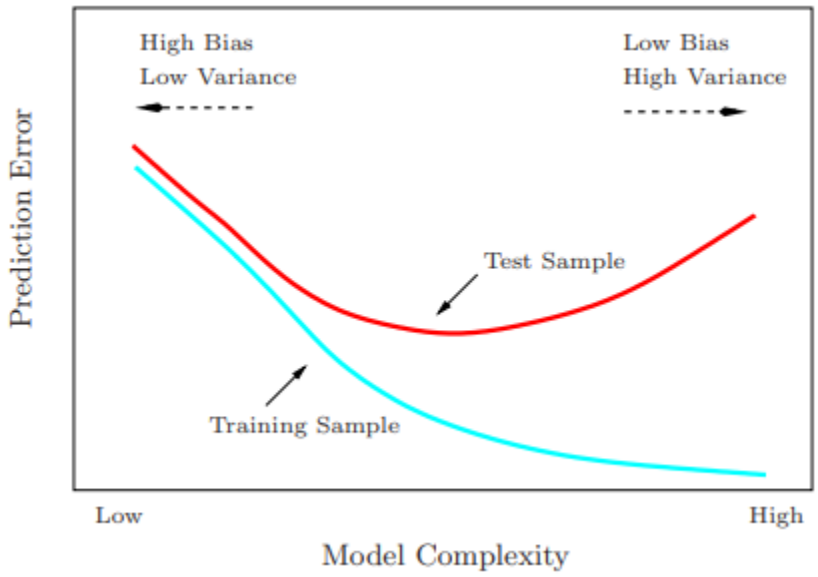


FIGURE 1.1. Bias-Variance Tradeoff, as discussed in Hastie et al. [2001].

Specifically, there is a built-in assumption that **interpolation** is inherently bad, i.e., if the training error is low, the test error must be high. The reasoning behind this belief is that here is a fundamental bias-variance tradeoff: Models of low complexity should can have low variance but high bias because, whereas models with high complexity may have low bias but must have high variance because they interpolate the data. As such at either extreme one should have poor generalization and you need to judiciously select the sweet spot in between. Even at this level of discussion, however, the story is not so transparent. Consider the following classical example of k-Nearest Neighbors regression.

EXAMPLE 5.1 (k-Nearest Neighbors regression). Consider the data described by $Y = f(X) + \varepsilon$, where $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$. We'll use the k-nearest neighbor algorithm to estimate the function:

$$\hat{f}_k(x) = \frac{1}{k} \sum_{\ell} f(x_{(\ell)})$$

This sums over the k-nearest points to x. The test error will be

$$\begin{aligned}\text{Test}(x_0) &= \mathbb{E}[(Y - \hat{f}_k(x_0))^2 \mid X = x_0] \\ &= \sigma^2 + \text{Bias}^2(\hat{f}_k) + \text{Var}(\hat{f}_k) \\ &= \sigma^2 + \left(f(x_0) - \frac{1}{k}\sum f(x_{(\ell)})\right)^2 + \frac{\sigma^2}{k}\end{aligned}$$

However, this does **not** fully align with what the diagram above suggests.

Today, we are going to take a deeper look at this story. We'll see that, despite the clear intuitive appeal, the popular wisdom

$$\text{Low Training Error} \neq \text{High Test Error}$$

is incorrect.

2. Non-Parametric Regression (Kernel Smoothing)

Given data points $\{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ drawn independently and identically distributed (iid), our objective is to estimate the posterior mean function: $f(x) = \mathbb{E}[Y \mid X = x]$. Our goal is to construct a **non-parametric estimator** for $f(x)$, where here we will use the **Nadaraya-Watson estimator ('64)** to estimate $f(x)$:

$$f_n(x) = \begin{cases} Y_i & \text{if } x = X_i \text{ for some } i \in [n], \\ 0 & \text{if } \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) = 0, \\ \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)} & \text{otherwise} \end{cases},$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a **kernel function**, and $h > 0$ is the **bandwidth parameter**. The kernel function assigns weights to the data points, with more weight assigned to points closer to x, and the bandwidth controls the smoothness of the estimator.

The idea is that we're going to smooth the concept of k-nearest neighbor by replacing indicators with kernels. Here, $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel and $h > 0$ is a fixed *bandwidth* parameter. For simplicity, we will write $K_h(x) = K\left(\frac{x}{h}\right)$. Our choice of kernel will be

$$K(u) = \frac{1}{\|u\|_2^a} \mathbb{1}(\|u\|_2 \leq 1),$$

where $0 < a < \frac{d}{2}$.

The average distance between any n points in a unit cube of dimension d should be approximately $n^{-\frac{1}{d}}$ apart. If the bandwidth h is small enough, only one of the points in the kernel will be nonzero. Therefore, for $h \asymp n^{-1/d}$, f_n will interpolate the data. We will now show that, under certain conditions, this interpolation will also lead to low test error.

Definition 5.2. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called **β-Hölder continuous** if

$$|f(x) - f(y)| \leq C_f \|x - y\|^\beta \quad (*)$$

The smallest such C is called the **β-Hölder (quasi)norm** of f, i.e.

$$\|f\|_\beta = \text{smallest } C_f \text{ such that } (*) \text{ holds.}$$

We will consider $B(\beta, R) = \{f : \|f\|_\beta \leq R\}$. Rather than imposing a parametric assumption on the function class, we will make a base regularity assumption about the function. We're going to work under the following assumption: recall that our data is of the form $Y = f(X) + \varepsilon$.

A1. f exists and $\mathbb{E}[(Y - f(X))^2 \mid X] \leq \sigma_\varepsilon^2 < \infty$
A2. On $\text{supp}(P_X)$, X has density $0 \leq p_{\min} < p(x) \leq p_{\max} < \infty$

THEOREM 5.3 (Belkin et al. [2019]). *Suppose that assumptions A1 and A2 hold, $\|f\|_\beta \leq L$ for some $\beta \in (0, 1]$ and $L > 0$. We have $0 < a < \frac{d}{2}$. Under these assumptions, then $\forall x_0 \in supp(P_x) \subseteq \mathbb{R}^d$, if $h = n^{-\frac{1}{d+2\beta}}$ then*

$$\mathbb{E}[(f(x_0) - f_n(x_0))^2] \lesssim n^{-\frac{2\beta}{2\beta+d}}.$$

Furthermore, the same inequality holds for the expected test error:

$$\mathbb{E}[(f(X_0) - f_n(X_0))^2] \lesssim n^{-\frac{2\beta}{2\beta+d}}.$$

PROOF. WLOG, let $x_0 = 0$, $|f(x_0)| < \infty$. Let

$$E = \left\{ \sum_{i=1}^n K(x_i) \neq 0 \right\} = \{\exists i : \|X_i\| \leq h\}.$$

We have

$$\begin{aligned}
\mathbb{P}(E^c) &\leq (1 - \mathbb{P}(\|X\| \leq h))^n \\
&\leq (1 - Cp_{\min}h^d)^n \leq \exp(-Cp_{\min}nh^d).
\end{aligned}$$

On the event E^c, we have $f_n(0) = 0$, so the contribution to the expected risk is at most:

$$f(0)^2\mathbb{P}(E^c) \leq f(0)^2 \exp\left(-Cp_{\min}nh^d\right).$$

Thus, the overall bound for the expected squared error is:

$$\mathbb{E}[(f(0) - f_n(0))^2] \leq \mathbb{E}[(f(0) - f_n(0))^2 \mid E] + f(0)^2 \exp\left(-Cp_{\min}nh^d\right).$$

Now, we will bound the first term by decomposing it into bias and variance terms:

$$\mathbb{E}[(f(0) - f_n(0))^2, E] = b^2(0) + V(0),$$

where $b(0)$ represents the bias and $V(0)$ represents the variance. The bias term is given by:

$$\begin{aligned}
b^2(0) &= \mathbb{E}\left[\left(\mathbb{E}_{Y|X} f_n(0) - f(0)\right)^2, E\right] \\
&= \mathbb{E}\left[\left(\frac{\sum_i (f(X_i) - f(0))K_h(X_i)}{\sum_i K_h(X_i)}\right)^2, E\right].
\end{aligned}$$

Recall that f is Hölder continuous with $\|f\|_\beta \leq L$. Therefore, using the bound $\|f(X_i) - f(0)\| \leq L\|X_i\|^\beta$ for all i, we get:

$$\left(\frac{\sum_i (f(X_i) - f(0))K_h(X_i)}{\sum_i K_h(X_i)}\right)^2 \leq L^2 \left\|\frac{\sum_i \|X_i\|^\beta K_h(X_i)}{\sum_i K_h(X_i)}\right\|^2 \leq L^2 h^{2\beta}$$

where the last inequality holds since only points X_i such that $\|X_i\| \leq h$ contribute to the kernel summation. Thus, the entire expectation for the bias term as follows: $b^2(0) \leq L^2 h^{2\beta}$. Let us now bound the variance. Let $\varepsilon_i = Y_i - f(X_i)$.

$$\begin{aligned}
V(0) &= \mathbb{E}\left[\left(\frac{\sum_i \varepsilon_i K_h(X_i)}{\sum_i K_h(X_i)}\right)^2, E\right] \\
&= \mathbb{E}\left[\frac{\sum_i \varepsilon_i^2 K_h(X_i)^2}{(\sum_i K_h(X_i))^2}, E\right] + \underbrace{\mathbb{E}\left[\frac{\sum_i \varepsilon_i \varepsilon_j K_h(X_i) K_h(X_j)}{(\sum_i K_h(X_i))^2}, E\right]}_{0 \quad \text{as } \mathbb{E}[\varepsilon_i]=0} \\
&\leq \sigma_\varepsilon^2 \cdot \mathbb{E}\left[\frac{\sum_i K_h(X_i)^2}{(\sum_i K_h(X_i))^2}, E\right] \\
&= n\sigma_\varepsilon^2 \cdot \mathbb{E}\left[\frac{K_h(X_1)^2}{(\sum_i K_h(X_i))^2}, E\right]
\end{aligned}$$

We will now bound $\mathbb{E}\left[\frac{K_h(X_1)^2}{(\sum_i K_h(X_i))^2}, E\right]$.

To do so, we define $\eta_i = \mathbb{1}(\|X_i\| \leq h) \overset{iid}{\sim} \text{Ber}(\overline{p})$ where $\overline{p} = \mathbb{P}(\|X\| \leq h)$. For $h < 1$, $K_h(X_i) \geq \eta_i$.

Therefore, we have

$$\mathbb{E}\left[\frac{K_h(X_1)^2}{(\sum_i K_h(X_i))^2}, E\right] \leq \underbrace{\mathbb{E}\left[\frac{K_h(X_1)^2}{(\sum_i K_h(X_i))^2}, \frac{1}{n}\sum_{i=1}^n \eta_i \leq \frac{1}{2}\overline{p}, E\right]}_{(A)} + \underbrace{\frac{4}{(n\overline{p})^2}\mathbb{E}\left[K_h(X_1)^2\right]}_{(B)}$$

We can bound (B) as follow:

$$\begin{aligned}
(B) &= \frac{4}{(n\overline{p})^2}\mathbb{E}\left[K\left(\frac{X_1}{h}\right)^2\right] && \text{using substitution } \frac{x}{h} = u \\
&= \frac{4}{(n\overline{p})^2} h^d \int_{\|u\|_2 \leq 1} K(u)^2 p(u)\, du \\
&\leq \frac{4}{(n\overline{p})^2} h^d p_{\max} \int_{\|u\|_2 \leq 1} K(u)^2\, du
\end{aligned}$$

Since the kernel K is radially symmetric and supported on the unit Euclidean ball, we can bound this as

$$\frac{4}{(n\overline{p})^2} h^d p_{\max} \int_{\|u\|_2 \leq 1} K(u)^2\, du \leq \frac{4}{(n\overline{p})^2} h^d p_{\max} \int_0^1 r^{-2a} r^{d-1}\, dr \lesssim \frac{h^d}{(n\overline{p})^2}$$

for some $a < \frac{d}{2}$. It thus follows that $(B) \lesssim \frac{h^d}{(n\overline{p})^2}$. As $\overline{p} = \mathbb{P}(\|X\| \leq h) \geq C_d p_{\min} h^d$. This therefore gives us $(B) \lesssim \frac{1}{n^2 h^d}$.

For term (A),

$$\begin{aligned}\mathbb{E}\left[\frac{K_h(X_i)^2}{(\sum_i K_h(X_i))^2), \frac{1}{n}\sum_{i=1}^n \eta_i \le \frac{1}{2}\overline{p}, E\right] &\le \mathbb{P}\left(\frac{1}{n}\sum_i \eta_i \le \frac{1}{2}\overline{p}\right) \\ &\le \mathbb{P}\left(\left|\frac{1}{n}\sum_i \eta_i - \overline{p}\right| \ge \frac{1}{2}\overline{p}\right) \\ &\overset{(*)}{\lesssim} \exp\left(-2n\left(\frac{\overline{p}}{2}\right)^2\right) \\ &\lesssim \exp\left(-\frac{n}{2}h^{2d}\right)\end{aligned}$$

where the inequality $(*)$ holds because of Hoeffding's inequality applied on Bernoulli random variables $\eta_1, \ldots, \eta_n \in [0, 1]$.

Wrapping up, as $d \ge 2$ and plugging in our choice of bandwidth $h = n^{-\frac{1}{2\beta+d}}$, we have that our test error can be bounded as

$$\text{Test}(0) \lesssim e^{-cnh^d} + L^2h^{2\beta} + \frac{1}{nh^d} + n\exp(-Cnh^{2d}) \lesssim n^{-\frac{2\beta}{2\beta+d}}$$

for some constants c and C, which proves our result. □

3. Double Descent

We should never assume that achieving zero training error necessarily implies a large generalization error. What happens to this U-shaped curve? This leads us to a phenomenon called as **double descent**. The moral story for double descent is that: what actually happens in various problems is that the picture that emerges will actually be as Figure 3.1.

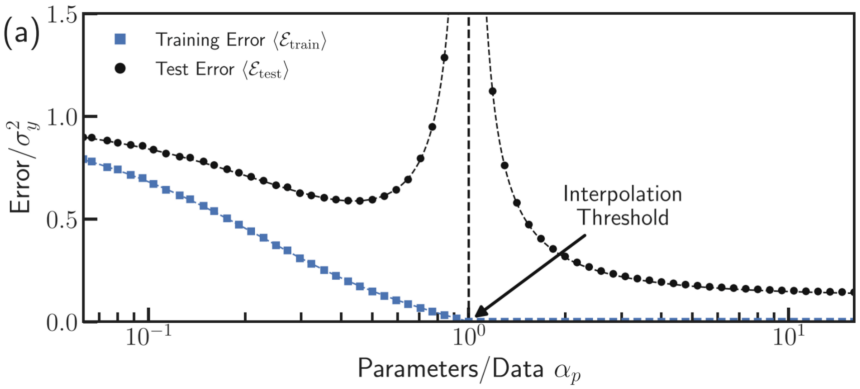


FIGURE 3.1. Double Descent, as in Schaeffer et al. [2023]

The simplest example to illustrate this phenomenon is **ridgeless least squares**. We assume our data (Y_i, X_i) are i.i.d., following the linear model: $y_i = \langle X_i, \beta \rangle + \varepsilon_i$ where ε_i here represents independent noise. The covariates $X_i \in \mathbb{R}^p \sim \mathcal{N}(0, I)$ and we are particularly interested in the high-dimensional regime, where the number of features p is much larger than n. In fact, we're going to assume that the ratio $\frac{p}{n}$ converges to some parameter γ, i.e.

$\frac{p}{n} \to \gamma$.

Let $\hat{\beta}$ be the minimum ℓ_2-norm least squares estimate. We would like to compute $R(\hat{\beta})$. There are two distinct regimes:

- **Underparameterized Regime**, corresponds to $\gamma \in (0, 1)$.

$$R(\hat{\beta}) \to \frac{\sigma_\varepsilon^2 \gamma}{1-\gamma} \quad \text{a.s.}$$

- **Overparameterized Regime**, corresponds to $\gamma \in (1, \infty)$.

$$R(\hat{\beta}) \to r^2 \left(1 - \frac{1}{\gamma}\right) + \sigma_\varepsilon^2 \left(\frac{1}{\gamma - 1}\right)$$

  where $\|\beta\| = r$.

4. Appendix

THEOREM 5.4 (Hoeffding's Inequality). *Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables with $\mathbb{P}(a \leq Z_i \leq b) = 1$ for all i and common mean μ, then for any $t > 0$,*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| > \varepsilon\right] \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}.$$

CHAPTER 6

A Primer on Random Matrix Theory

1. Introduction

To prove many of the results in the later parts of the course, we require key results from random matrix theory. Random matrix theory is the study of random matrices, often as they become large and will provide insights to our understanding of novel high-dimensional phenomena. We will focus on the following running example to motivate the ideas of this subject.

EXAMPLE 6.1 (Covariance Matrix Estimation). Throughout for this lecture we assume that $X_1, \ldots, X_M$ be IID $\mathbb{R}^N$-valued random vectors with X_i having IID entries and satisfying the following conditions:

- $\mathbb{E}X_i(k) = 0$
- $\sup_{k\in[N]} \mathbb{E}|X_i(k)|^4 < \infty$
- $\mathbb{E}X_i X_i^T = I$

where $X_i(k)$ denotes the k^{th} entry of X_i. Let $\hat{\Sigma}_M = \frac{1}{M}\sum X_k X_k^T$ denote the empirical covariance matrix. Then, the natural question is: will $\hat{\Sigma}_M \to I$ in some mode of convergence?

Classical asymptotic theory states that for N fixed we have that $\hat{\Sigma}_M \to I$ almost surely in any desired norm. However, this result essentially requires $M = \infty$ and is unrealistic for our high-dimensional setting.

EXAMPLE 6.2 (CIFAR-10 Dataset). In this image classification dataset, there are typically $M \sim 5 \times 10^4$ samples and each X_i is the vectorization of an image with $N \sim 10^3$ components. Therefore, it is unrealistic to suppose that $M >> N$.

For this lecture, suppose we are in the regime where M and N scale together. Let $\alpha_N = \frac{M}{N}$. If $\alpha_N >> 1$ and we suppose the necessary moment conditions, then $\|\hat{\Sigma}_M - I\|_{\text{op}}$ will be small. However, in the limit, as $\alpha_N \to \alpha \in (0, \infty)$ it can be shown that $\underline{\lim}\, \|\hat{\Sigma}_M - I\|_{\text{op}} > c > 0$.

What notions of convergence can we have for the covariance matrix? From the perspective of RMT we will study the spectral properties of $\hat{\Sigma}_M$.

DEFINITION 6.3 (Wishart Distribution). $\hat{\Sigma}_M$ is called a Wishart distribution.

Denote $W_N := \hat{\Sigma}_M$ where N denotes the dimension of W_N. Since W_N has dimension N and is a real symmetric matrix, this implies that W_N has N real eigenvalues. Denoting the ordered eigenvalues as $\lambda_1^N, \lambda_2^N, \ldots, \lambda_N^N$ we have for the empirical spectral measure as

DEFINITION 6.4 (Empirical Spectral Measure). The empirical spectral measure denoted as μ_N is defined by

$$\mu_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i^N}.$$

2. Marchenko-Pastur Law

The idea of this section is to show that the (weak) limiting distribution of μ_N is the following distribution

DEFINITION 6.5 (Marchenko-Pastur Distribution). The Marchenko-Pastur distribution with parameters σ and α is the measure $\mu_{\sigma,\alpha}$ with

$$\mu_{\alpha,\sigma}(A) = \begin{cases} \left(1-\frac{1}{\alpha}\right)\mathbf{1}_{0\in A} + \nu_{\sigma,\alpha}(A), & \text{if } \alpha > 1 \\ \nu_{\sigma,\alpha}(A), & \text{if } 0 \le \alpha \le 1, \end{cases}$$

and $\nu_{\sigma,\alpha}$ has density given by

$$d\nu(x) = \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\lambda+ - x)(x-\lambda_-)}}{\alpha x}\mathbf{1}_{x\in[\lambda_-,\lambda_+]}dx$$

with

$$\lambda_\pm = \sigma^2(1\pm\sqrt{\lambda})^2.$$

The Marchenko-Pastur law gives the following convergence for the spectral measure.

THEOREM 6.6 (Marčenko and Pastur [1967]). *Assume that $M, N \to \infty$ with limiting ratio $\frac{M}{N} \to \alpha \in (0,\infty)$. Then, $\mu_N \to \mu_{\alpha,\sigma}$ weakly almost surely. That is,*

$$P(\mu_N \to \mu_{\alpha,\sigma} \text{ weakly }) = 1.$$

Before we can prove this theorem, we need to recall some notions from spectral calculus.

DEFINITION 6.7 (Resolvent). If A is an $N\times N$ Hermitian matrix, the Resolvant $R_A(z)$ is defined by

$$R_A(z) = (A-z)^{-1}.$$

REMARK 6.8. R_A is a matrix valued meromorphic function (holomorphic except at isolated singularities) with poles at the spectrum of A.

THEOREM 6.9 (Cauchy Integral Formula for Matrices). *Suppose that $f : \mathbb{C}\to\mathbb{C}$ is analytic and Γ is the contour that encloses all eigenvalues of the matrix A. Then, we have the following representation*

$$f(A) = \frac{1}{2\pi i}\int_\Gamma \frac{f(z)}{A-z}dz.$$

REMARK 6.10. Note that since the matrix A is Hermitian we will also have that

$$f(A) = \sum_i f(\lambda_i)v_i v_i^T.$$

Additionally, since $\operatorname{tr}\left(f(\lambda_i)v_i v_i^T\right) = f(\lambda_i)$ we see that

$$\frac{1}{N}\operatorname{tr}(f(A)) = \frac{1}{N}\sum_i f(\lambda_i) = \int f d\mu_N.$$

DEFINITION 6.11 (Stieltjes Transform). Let μ be a probability measure on $\mathbb{R}$ the Stieltjes transform of μ is given by

$$S_\mu(z) = \int \frac{d\mu(x)}{x - z} \quad z \in \mathbb{C} \setminus \mathbb{R}.$$

REMARK 6.12. Sometimes we write S_{W_N} or even S_N in place of S_{μ_N} where μ_N is the spectral measure of W_N. In all cases we understand $S_{(\cdot)}$ as the Stieltjes transform of the spectral measure of a Wishart distribution.

From the definition it is clear that S_μ is analytic on $\mathbb{C} \setminus \mathbb{R}$. In particular, we have the following relationship between the resolvent and the Stieltjes transform:

$$\operatorname{tr}(\operatorname{Res}_{W_N}(z)) = S_{\mu_N}(z).$$

Moreover, the Stieltjes transform S_μ uniquely determines the measure μ and there is an inversion formula that allows the recovery of μ from S_μ.

THEOREM 6.13 (Stieltjes Inversion). *For any interval $I = (a, b] \subset \mathbb{R}$, $\mu(\{a\}) = \mu(\{b\}) = 0$*

$$\mu(I) = \lim_{\epsilon \to 0^+} \frac{1}{\pi} \int_I \operatorname{Im} S_\mu(\lambda + i\epsilon)\, d\lambda.$$

PROOF SKETCH. The idea is to show that for $X \sim \mu$, $C_\epsilon \sim \text{Cauchy}(\epsilon)$ then $\frac{1}{\pi}\operatorname{Im}(S_\mu(\lambda + i\epsilon))d\mu$ is the law of $X + C_\epsilon$. Recall that the pdf of the Cauchy(ϵ) distribution is given by

$$dC_\epsilon(x) = \frac{\epsilon}{\pi(x^2 + \epsilon^2)}.$$

Since $\operatorname{Im}(\frac{1}{z}) = -\frac{\operatorname{Im}(z)}{|z|^2}$ this implies that

$$\frac{1}{\pi}\operatorname{Im}(S_\mu(\lambda + i\epsilon))d\mu = \frac{1}{\pi}\int \frac{\epsilon}{(z - x)^2 + \epsilon^2} d\mu(x)$$

This is the convolution of the pdfs of X and C_ϵ which proves the claim. Now since the pdf of C_ϵ approaches a dirac delta as $\epsilon \to 0$ take the limit in the appropriate sense and the result follows. □

The inversion theorem allows us to establish the following result. Consequently to prove the weakly almost sure convergence we will establish convergence of the Stieltjes transform.

THEOREM 6.14. *Let $\mu_n \in \mathscr{M}_1(\mathbb{R})$ (the space of probability measures on $\mathbb{R}$). Then the following statements hold:*

- *$\mu_n \to \mu$ vaguely iff $S_{\mu_n} \to S_\mu$ on $\mathbb{C} \setminus \mathbb{R}$ pointwisely.*
- *For μ random and μ deterministic, $\mu_n \to \mu$ weakly almost surely iff $S_{\mu_n}(z) \to S_\mu(z)$ $\forall z \in \mathbb{C} \setminus \mathbb{R}$.*

3. Proof of the Marchenko–Pastur Law

There are 3 ingredients to the proof:

(1) Self consistency equation for the Stieltjes transform,
(2) Leave-one-out identity,
(3) Concentration bound.

3.1. Self Consistency. From now on σ, α are fixed and we write S in the place of S_μ for the transform of the Marchenko-Pastur law. Through algebra it can be shown that S satisfies the self consistency equation:

$$1 + zS(z) = \frac{S(z)}{1 + \frac{1}{\alpha}S(z)}.$$

3.2. Sherman-Morrison Identity.

THEOREM 6.15 (Sheman-Morrison). *Suppose that A is invertible, then A satisfies*

$$(A + tvvT)^{-1}v = \frac{A^{-1}v}{1 + t\langle v, Av\rangle}$$

3.3. Concentration Argument. Before we prove Theorem 6.6 we need two kinds of results:

- $S_N(z) = S_{W_N}(z)$ has almost sure limit
$$S_N(z) \to \mathbb{E}S_N(z) \to 0$$
- $\frac{1}{N}\langle X_N, W_{N-1}X_N\rangle - S_N(z) \to 0$ almost surely

LEMMA 6.16. *For the Stieltjes transform of Wishart matrices,*

$$S_N(z) \to \mathbb{E}S_N(z) \quad \textit{almost surely}$$

In fact, if $|Im(z)| = v > 0$ then

$$P(|S_N(z) - \mathbb{E}S_N(z)| > t) \lesssim e^{\frac{-Nct^2}{|Im(z)|}}.$$

PROOF. The idea of the proof is to decompose $S_N(z) - \mathbb{E}S_N(z)$ as a bounded martingale difference. Let

$$\mathcal{F}_k = \sigma(X_{k+1}, \ldots, X_M) \text{ and } \mathcal{F}_M \text{ be the trivial } \sigma \text{ algebra}$$

and $\mathbb{E}_l := \mathbb{E}(\cdot \mid \mathcal{F}_l)$. Moreover, let $W_N^\ell = W_N - \frac{1}{M}X_\ell X_\ell^T$. Since W_N^ℓ is independent of X_ℓ this implies that $\mathbb{E}_\ell S_{W_N^\ell}(z) = \mathbb{E}_{\ell-1}S_{W_N^\ell}(z)$. Now from the above we write $S_N(z) - \mathbb{E}[S_N(z)]$ as a sum of martingale differences,

$$\begin{aligned} S_N(z) - \mathbb{E}S_N(z) &= \sum_{k=1}^{M} \mathbb{E}_{k-1}S_N(z) - \mathbb{E}_k S_N(z) \\ &= \sum_{k=1}^{M} \mathbb{E}_{k-1}S_N(z) + (\mathbb{E}_k S_{W_N^k} - \mathbb{E}_{k-1}S_{W_N^k}) - \mathbb{E}S_N(z) \\ &= \sum_{k=1}^{M} (\mathbb{E}_{k-1} - \mathbb{E}_k)(S_N(z) - S_{W_N^k}(z)) =: \sum_{k=1}^{M} \gamma_k. \end{aligned} \tag{1}$$

Observe that for real symmetric matrices A, with $\mathrm{Im}(z) > 0$ and $c \in \mathbb{R}$ we have that

$$(A + cvv^T - zI)^{-1} - (A - zI)^{-1} = -c(A - zI)^{-1}vv^T(A + cvv^T - zI)^{-1}.$$

This follows from an application of the matrix identity $X^{-1} - Y^{-1} = -Y^{-1}(Y - X)X^{-1}$. Thus, for any vector $v \in \mathbb{R}^N$,

$$\begin{aligned}\operatorname{tr}\left((A + cvv^T - zI)^{-1} - (A - zI)^{-1}\right) &= \operatorname{tr}\left((A + cvv^T - zI)^{-1} - (A - zI)^{-1}\right) \\ &= -c\frac{\operatorname{tr}\left((A - zI)^{-1}vv^T(A - zI)^{-1}\right)}{1 + c\langle v, (A - z)^{-1}v\rangle} \\ &= -c\frac{\langle v, R_A(z)^2 v\rangle}{1 + c\langle v, R_A(z)v\rangle}.\end{aligned}$$

Therefore, since the Stieltjes distribution is related to trace of the resolvent we apply this identity with $A = W_N^k, v = X_k, c = \frac{1}{M}$ and we get that

$$\begin{aligned}|S_N - S_{W_N^k}| &= \frac{1}{M}\left|\frac{\frac{1}{M}\langle X_k, R_{W_N^k}(z)^2 X_k\rangle}{1 + \frac{1}{M}\langle X_k, R_{W_N^k}(z)X_k\rangle}\right| \\ &\leq \frac{1}{M^2}\frac{\langle X_k, ((W_N^k - u)^2 + v^2 I)^{-1}X_k\rangle}{\mathrm{Im}(1 + \frac{1}{M}\langle X_k, R_{W_N^k}(z)X_k\rangle)} \leq \frac{1}{vM}\end{aligned}$$

since the imaginary part is less than the modulus and here $u = |\operatorname{Re}(z)|, v = |\operatorname{Im}(z)|$. In particular this shows that $S_N(z) - \mathbb{E}[S_N(z)]$ is a sum of bounded martingale differences and the results follows from a straight forward application of the Azuma inequality (Theorem 6.20). □

Remark 6.17. In the proof we denoted $W_N^l = W_N - \frac{1}{M}X_l X_l^T$. This notation will be used in the proof of Theorem 6.6.

Lemma 6.18. *Let (A_N) be a sequence of $N \times N$ matrices with $\|A_N\|_{\mathrm{op}} \leq C$, then*

$$\frac{1}{N}\langle X, AX\rangle - \frac{1}{N}\operatorname{tr}(A) \to 0 \quad \textit{in } L^1.$$

Proof. Let $D_N = \operatorname{diag}((A_N)_{ii})$. Note by Jensen's inequality

$$\mathbb{E}|\langle X, (A_N - D_N)X\rangle| \leq \sqrt{\mathbb{E}|\langle X, (A_N - D_N)X\rangle|^2},$$

and since the components of X, X_i, are mean zero, unit variance and independent,

$$\mathbb{E}|\langle X, (A_N - D_N)X\rangle|^2 = \mathbb{E}|\sum_{i\neq j} X_i(A_{ij})X_j|^2 = \sum_{i\neq j} A_{ij}^2 \leq N\|A\|_{\mathrm{op}}^2$$

Thus it follows that $\frac{1}{N}\left(\langle X, AX\rangle - \langle X, DX\rangle\right) \to 0$ in L^1. It can be checked that $\sum_i (X_i^2 - 1)$ is a martingale and since $|A_{ii}| \le \|A\|_{\mathrm{op}}$ we get by Burkholder's and Jensen's inequality that

$$\begin{aligned}
\mathbb{E}\left|\frac{\langle X, DX\rangle - \operatorname{tr}(A)}{N}\right| &= \mathbb{E}\left|\frac{\sum A_{ii}(X_i^2 - 1)}{N}\right| \\
&\le \frac{\|A\|_{\mathrm{op}}}{N}\mathbb{E}\left|\sum (X_i^2 - 1)\right| \\
\text{(Jensen's)} \qquad &\le \frac{\|A\|_{\mathrm{op}}}{N}\sqrt{\mathbb{E}\left(\sum (X_i^2 - 1)\right)^2} \\
\text{(Burkholder's)} \qquad &\le \frac{C\|A\|_{\mathrm{op}}}{N}\sqrt{\mathbb{E}\sum_i (X_i^2 - 1)^2} \\
&\lesssim \frac{1}{\sqrt{N}} \to 0.
\end{aligned}$$

where the last inequality follows from the moment assumptions. Combining the two inequalities gives the result. □

Now we are in a position to prove Theorem 6.6.

Proof of Theorem 6.6. By definition,

$$R_W(z)(W - z) = I.$$

Taking the trace and dividing by N gives

$$1 = \frac{1}{N}\operatorname{tr}(R_W(z)W) - zS_N(z).$$

Since $1 + zS_N(z)$ also concentrates about its mean by Lemma 1 we have that

$$\frac{1}{N}Tr(R_w(z)W) - \mathbb{E}(\frac{1}{N}Tr(R_w(z)W)) \to 0 \text{ almost surely .} \tag{2}$$

Recall that $R_W W = \frac{1}{M}\sum R_W(z)X_k X_k^T$ and so by Sherman-Morrison

$$R_W X_k = \frac{R_{W_N^k} X_k}{1 + \frac{1}{M}\langle X_k, R_{W_N^k} X_k\rangle}.$$

Now taking expectations,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N}\operatorname{tr}(R_W W)\right] &= \frac{1}{N}\mathbb{E}\left[\frac{1}{M}\sum_k \frac{\operatorname{tr}(R_{W_N^k} X_k X_k^T)}{1 + \frac{1}{M}\langle X_k, R_{W_N^k} X_k\rangle}\right] \\
&= \frac{1}{N}\mathbb{E}\left[\frac{1}{M}\sum_k \frac{\langle X_k, R_{W_N^k} X_k\rangle}{1 + \frac{1}{M}\langle X_k, R_{W_N^k} X_k\rangle}\right] \\
&= \mathbb{E}\left[\frac{\frac{1}{N}\langle X_k, R_{W_N^N} X_k\rangle}{1 + \frac{1}{M}\langle X_k, R_{W_N^N} X_k\rangle}\right]
\end{aligned}$$

where we used the fact that the distribution of each term in the sum is the same for all k. Since $\mathrm{tr}(\mathrm{Res}_{W_N}) = S_N$ and $\frac{N}{M} \to \frac{1}{\alpha}$, by Lemma 2 as $\|R_{W_N}\|_{\mathrm{op}} \leq \frac{1}{|\mathrm{Im}(z)|}$ we have for each $z \in \mathbb{C} \setminus \mathbb{R}$ that

$$\mathbb{E}\left[\frac{\frac{1}{N}\langle X_k, R_{W_N^N}(z)X_k\rangle}{1+\frac{1}{M}\langle X_k, R_{W_N^N}(z)X_k\rangle}\right] \to \frac{\lim S_N(z)}{1+\frac{1}{\alpha}\lim S_N(z)}.$$

Therefore, by the bounded convergence theorem this implies that

$$\mathbb{E}\left[\frac{\frac{1}{N}\langle X_k, R_{W_N^N}X_k\rangle}{1+\frac{1}{M}\langle X_k, R_{W_N^N}X_k\rangle}\right] \to \frac{\lim S_N}{1+\frac{1}{\alpha}\lim S_N} \text{ on } \mathbb{C} \setminus \mathbb{R}.$$

Hence S_N converges to the solution of

$$1 + zS(z) = \frac{S(z)}{1+\frac{1}{\alpha}S(z)}.$$

which by Theorem 6.14 proves the Marchenko-Pastur distribution. □

4. Appendix

As a reminder these are the definitions of weak notions of convergence.

Definition 6.19. Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence in in $\mathcal{M}_1(\mathbb{R})$ and let $\mathcal{C}_c(\mathbb{R})$ denote the continuous functions with compact support and and $\mathcal{C}_b(\mathbb{R})$ denote the continuous and bounded functions.
i) The sequence $(\mu_n)_{n\in\mathbb{N}}$ is said to converge weakly to an element $\mu \in \mathcal{M}_1(\mathbb{R})$, if

$$\forall f \in \mathcal{C}_b(\mathbb{R}) : \lim_{n\to\infty} \int f d\mu_n = \int f d\mu$$

ii) The sequence $(\mu_n)_{n\in\mathbb{N}}$ is said to converge vaguely to an element $\mu \in \mathcal{M}_{\leq 1}(\mathbb{R})$, if

$$\forall f \in \mathcal{C}_c(\mathbb{R}) : \lim_{n\to\infty} \int f d\mu_n = \int f \mu$$

For the inequalities, recall that a martingale difference sequence is an adapted sequence $\{D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ such that for all $k \geq 1$,

$$\mathbb{E}[|D_k|] < \infty \quad \text{and} \quad \mathbb{E}[D_{k+1} \mid \mathcal{F}_k] = 0.$$

The following is a useful concentration inequality for sums of bounded martingale differences

Theorem 6.20 (Azuma's Inequality). *Let $\{D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ be a martingale difference sequence for which there are constants $\{a_k, b_k)\}_{k=1}^{n}$ such that $|D_k| \leq c_k$ almost surely for all $k = 1, \ldots, n$. Then, for all $t \geq 0$,*

$$P\left[\left|\sum_{k=1}^{n} D_k\right| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{2\sum_{k=1}^{n} c_k^2}\right)$$

For the sake of simplifying the proof of Lemma 2 we state the following simplified version of the Burkholder-Davis-Gundy inequality.

THEOREM 6.21 (Burkholder's Inequality). *Let* $\{(X_i, \mathcal{F}_i)\}_{i=1}^n$ *be a martingale. Define martingale indcrements* $\xi_1 = X_1$ *and* $\xi_i = X_i - X_{i-1}$ *for* $i = 2, \ldots, n$. *The quadratic variation process is defined as*

$$Q_i = \xi_1^2 + \ldots \xi_i^2 \quad \text{for } i = 1, \ldots, n$$

Burkholder's inequality shows that, as far as L^p *norms are concerned,* $\sqrt{Q_n}$ *and* X_n *increase at the same rate: for each* $p > 1$ *there exist positive constants* c_p *and* C_p *such that*

$$c_p \left\| \sqrt{Q_n} \right\|_p \leq \|X_n\|_p \leq C_p \left\| \sqrt{Q_n} \right\|_p .$$

CHAPTER 7

Double Descent

1. Ridgeless Least Squares

As was mentioned in passing in Lecture 5 we should have never assumed that interpolation implies bad test error.

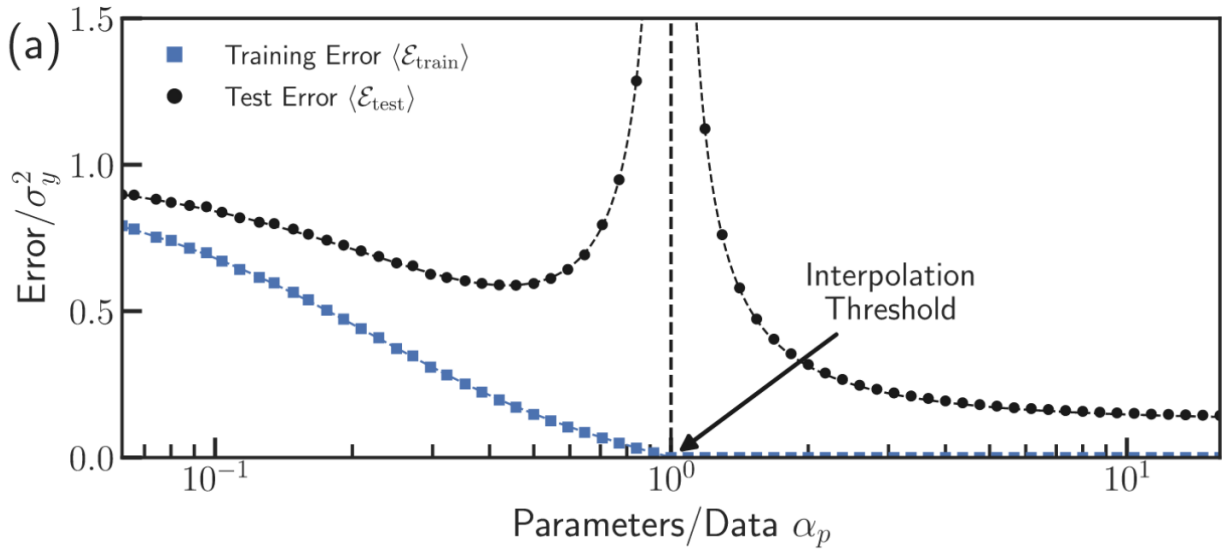


FIGURE 1.1. Schaeffer et al. [2023]

We want to show that the classical U-shape curve is a case of early confirmation bias. The picture for the double descent of ridgeless least squares is analogous to the one in Figure 1.1.

EXAMPLE 7.1 (Ridgless Least Squares). Suppose we are given data $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^p$ where $(x_i, \epsilon_i) \sim \mathcal{P}_x \otimes \mathcal{P}_\epsilon$ for $i = 1, \dots, n$ IID. The relation between the features and the target is given by

$$y_i = \langle x_i, \beta_i \rangle + \epsilon_i,$$

where $\mathbb{E}x_i = 0$, $\text{Cov}(x_i) = I$, $\mathbb{E}\epsilon_i = 0$, $\mathbb{E}\epsilon_i^2 = \sigma^2$. The goal in ridgeless regression is to infer β. Due to overparameterization there will be infinitely many solutions. Thus, letting $X = [x_1, \dots, x_n] \in \mathbb{R}^{n \times p}$ and $y = (y_1, \dots, y_n)^T$ the estimator $\hat{\beta}$ is the minimum l_2 estimator solving

$$\min \|\tilde{\beta}\| \tag{1.1}$$

$$\text{such that } \tilde{\beta} \in \arg\min \|y - X^T \beta\| \tag{1.2}$$

DEFINITION 7.2 (Moore-Penrose Pseudoinverse). For a matrix $A \in \mathbb{R}^{N\times M}$ its Moore-Penrose pseudoinverse is the matrix A^+ defined from its singular value decomposition $A = UDV^T$ by writing

$$D^+ = \operatorname{diag}\left(\frac{1}{D_{ii}}\mathbb{1}_{(D_{ii}\neq 0)}\right)$$

and setting

$$A^+ = VD^+U^T.$$

From this definition, it can be observed that

$$A^+A = V\operatorname{diag}(\mathbb{1}_{D_{ii}\neq 0})V^T,$$

with the analogous result for U in the place of V in AA^+ and the weak inverse property

$$A^+AA^+ = A^+.$$

fact* The solution to ridgeless least squares is given by

$$\hat{\beta} = (XX^T)^+Xy.$$

fact*

Additionally, $\hat{\beta}$ can be interpreted as the limit of gradient descent on the least square objective as was seen in a previous lecture. Now that we know the estimate $\hat{\beta}$ a natural question is to calculate the out-of-sample risk. Let X_0 be an independent copy of X, we obtain a similar bias variance decomposition. Indeed,

$$\begin{aligned}
R(\hat{\beta},\beta) &= \mathbb{E}\left[(\langle X_0, \hat{\beta}-\beta\rangle)^2 \mid X\right] \\
&= \mathbb{E}\left[(\hat{\beta}-\beta)^T X_0 X_0^T(\hat{\beta}-\beta) \mid X\right] \\
&= \operatorname{tr}\left(\mathbb{E}\left[X_0X_0^T\right]\mathbb{E}\left[\langle\hat{\beta}-\beta, \hat{\beta}-\beta\rangle \mid X\right]\right) \\
&= \mathbb{E}\left[\|\hat{\beta}-\beta\|_2^2 \mid X\right] \\
&= \mathbb{E}\left[\|(E(\hat{\beta}\mid X)-\beta) + (\hat{\beta}-E(\hat{\beta}\mid X))\|_2^2 \mid X\right] \\
&= \|\mathbb{E}(\hat{\beta}\mid X)-\beta\|^2 + \operatorname{tr}(\operatorname{Cov}(\hat{\beta}\mid X)) = B(\hat{\beta},\beta) + V(\hat{\beta},\beta)
\end{aligned}$$

where $B(\hat{\beta},\beta)$ is the bias and $V(\hat{\beta},\beta)$ is the variance. Note in the third line we used the fact that $\operatorname{Cov}(X_0) = I = \mathbb{E}[X_0X_0^T]$ and that X_0 is independent of X.

LEMMA 7.3. *Let $\hat{\Sigma}_n := \frac{XX^T}{n} \in \mathbb{R}^{p\times p}$ be the sample covariance matrix. Then denoting $\Pi := I - \hat{\Sigma}_n^+\hat{\Sigma}_n = \operatorname{Proj}(\operatorname{Ker}(X))$ we have that*

$$B_X(\hat{\beta},\beta) = \beta^T\Pi\Pi\beta = \beta^T\Pi\beta$$

$$V_X(\hat{\beta},\beta) = \frac{\sigma^2}{n}\operatorname{tr}(\hat{\Sigma}_n^+)$$

PROOF. From the fact it can be shown that $\mathbb{E}(\hat{\beta}\mid X) = \hat{\Sigma}_n^+\hat{\Sigma}_n\beta$. Moreover, as $A^+AA^+ = A^+$ it follows that $\operatorname{Cov}(\hat{\beta}\mid X) = \sigma^2(XX^T)^+XX^T(XX^T)^+ = \frac{\sigma}{n}\hat{\Sigma}_n^+$. □

2. Double Descent

Now we are in the position to state the main result for double descent.

THEOREM 7.4 (Double Descent in Ridgeless Regression). *Let X_i be IID with $\mathbb{E}X_i = 0, Cov(X_i) = 1, \mathbb{E}X_i^{4+\epsilon} < \infty$ for some $\epsilon > 0$. Let $r^2 = \|\beta\|_2^2$ for all n, p. If $\frac{p}{n} \to \gamma \in (0, \infty)$, then it holds almost surely that*

$$B(\hat{\beta}, \beta) \to r^2 \left(1 - \frac{1}{\gamma}\right),$$
$$V(\hat{\beta}, \beta) \to \sigma^2 \left(\frac{1}{\gamma - 1}\right).$$

Alternatively, if $\gamma \in (0, 1)$ then

$$B(\hat{\beta}, \beta) \to 0,$$
$$V(\hat{\beta}, \beta) \to \sigma^2 \left(\frac{\gamma}{\gamma - 1}\right).$$

The following theorem is the key result for the proof of Theorem 2.1

THEOREM 7.5 (Bai and Yin [1993]). *Ordering the eigenvalues of $\hat{\Sigma}_n$ as $\lambda_1 \leq \ldots \leq \lambda_n$, if $\frac{p}{n} \to \gamma < 1$ then it holds almost surely that*

$$\lambda_1(\hat{\Sigma}_n) \to (1 - \sqrt{\gamma})^2,$$
$$\lambda_1(\hat{\Sigma}_n) \geq (1 - \sqrt{\gamma})^2 \quad \textit{for n sufficiently large.}$$

Conversely, when $\gamma > 1$ it holds almost surely that

$$\lambda_{p-n+1}(\hat{\Sigma}_n) \to (1 - \sqrt{\gamma})^2,$$
$$\lambda_{p-n+1}(\hat{\Sigma}_n) \geq (1 - \sqrt{\gamma})^2 \quad \textit{for n sufficiently large.}$$

PROOF OF THEOREM 7.4. Suppose we have that $\lambda \in (0, 1)$. By Theorem 7.5, the matrix $\hat{\Sigma}_n$ is almost surely invertible and thus we have that $\Pi = I - \hat{\Sigma}_n^{-1}\hat{\Sigma}_n = 0$ almost surely which implies that $B_X(\hat{\beta}, \beta) = 0$ almost surely. Now denoting $d\mu_n$ as the empirical spectral measure of $\hat{\Sigma}_n$,

$$\begin{aligned} V_X(\hat{\beta}, \beta) &= \frac{\sigma^2}{n} \operatorname{tr}(\hat{\Sigma}_n^{-1}) \\ &= \frac{\sigma^2 \gamma}{p} \sum_{k=1}^{p} \frac{1}{\lambda_k} \\ &= \sigma^2 \gamma \int \frac{1}{t} d\mu_n. \end{aligned}$$

Since the spectrum is bounded from zero due to Bai-Yin, for n sufficiently large, picking K large implies that

$$\sigma^2 \gamma \int \frac{1}{t} d\mu_n = \sigma^2 \gamma \int \frac{1}{t} \wedge K d\mu_n \quad \text{for n large.}$$

By Marchenko Pastur as $\frac{1}{t} \wedge K$ is continuous and bounded,

$$V_X(\hat{\beta}, \beta) \to \sigma^2 \gamma \int \frac{d\mu(t)}{t}.$$

The integral on the right hand side is the Stieltjes transform of the Marchenko Pastur law evaluated at $z = 0$. From the previous lecture we saw that the self-consistency equation satisfies a quadratic equation and with some algebra it can be shown that

$$S_\gamma(-z) = \frac{-(1-\gamma+z) + \sqrt{(1-\gamma+z)^2 + 4\gamma z}}{2\gamma z}. \tag{1}$$

Therefore, taking the limit as $z \to 0^+$ in this expression yields $S_\gamma(0) = \frac{1}{1-\gamma}$ and the result follows.

Now we suppose that $\gamma > 1$. Recall Sylvester's determinant identity that for $A \in \mathbb{R}^{m \times m}$, $B \in \mathbb{R}^{n \times n}$,

$$\det(I_m + AB) = \det(I_n + BA),$$

which implies that the non-zero eigenvalues of $\frac{XX^T}{n}$ and $\frac{X^TX}{n}$ coincide. We will now find the limit of the variance. Since $p > n$, $\frac{XX^T}{n}$ has $p - n$ zero eigenvalues. Hence, by considering the non-zero eigenvalues of $\frac{X^TX}{n}$ we get that

$$\begin{aligned} V_X(\hat{\beta}, \beta) &= \frac{\sigma^2}{n} \sum_{i=1}^{n} \frac{1}{\lambda_{p+n+i}(\hat{\Sigma}_n)} \\ &= \frac{\sigma^2}{n} \sum_{i=1}^{n} \frac{1}{\lambda_i\left(\frac{X^TX}{n}\right)} \\ &= \frac{\sigma^2}{p} \sum_{i=1}^{n} \frac{1}{\lambda_i\left(\frac{X^TX}{p}\right)} \\ &= \frac{\sigma^2 n}{p} \int \frac{1}{t} d\mu_{\frac{X^TX}{p}}. \end{aligned}$$

Therefore, by the same argument as in the previous case,

$$V_X(\hat{\beta}, \beta) \to \frac{\sigma^2}{\gamma} S_{\frac{1}{\gamma}}(0) = \frac{\sigma^2}{\gamma} \times \frac{\gamma}{\gamma - 1} = \frac{\sigma^2}{\gamma - 1}.$$

Now it is time to consider the limit of the bias term. Note the following key characterization of the pseudoinverse of a rectangular matrix A

$$(AA^T)^+ A = \lim_{\delta \to 0} (AA^T + \delta I)^{-1} A.$$

We apply this result with $A = X^T/\sqrt{n}$ to write the bias term as

$$\begin{aligned} B_X(\hat{\beta}, \beta) = \beta^T \Pi \beta &= \beta^T (I - \hat{\Sigma}_n^+ \hat{\Sigma}_n)\beta \\ &= \lim_{z \to 0^+} \langle \beta, (I - (\hat{\Sigma}_n + z)^{-1} \hat{\Sigma}_n)\beta \rangle \\ &= \lim_{z \to 0^+} z \langle \beta, (\hat{\Sigma}_n + z)^{-1} \beta \rangle \\ &= \lim_{z \to 0^+} z \langle \beta, R_{\hat{\Sigma}_n}(-z)\beta \rangle \end{aligned} \tag{2}$$

To proceed we have to make use of Theorem 1 in Rubio and Mestre [2011], which may be seen as a generalized Marchenko-Pastur theorem.

THEOREM 7.6 (Rubio-Mestre (Reduced Version)). *Denoting $r^2 = \|\beta\|$ and m as the Stieltjes transform of the Marchenko-Pastur law F_γ it holds almost surely that*

$$\langle \beta, R_{\hat{\Sigma}_n}(-z)\beta \rangle \to r^2 S_\gamma(-z).$$

from which the result for the bias follows.

The proof of this result is similar to that of Marchenko Pastur and a sketch of the proof will be given in the Appendix. Going back to the proof of Theorem 7.4, let

$$f_n(z) = z \langle \beta, R_{\hat{\Sigma}_n}(-z)\beta \rangle.$$

Notice that $|f_n(z)| \leq r^2$, and $f_n'(z) = \beta^T (\hat{\Sigma} + zI)^{-2} \hat{\Sigma} \beta$, so

$$|f_n'(z)| \leq r^2 \frac{\lambda_{\max}(\hat{\Sigma})}{\left(\lambda_{\min}^+(\hat{\Sigma}) + z\right)^2} \leq C \quad \text{almost surely.}$$

where $\lambda_{\max}(\hat{\Sigma}_n)$ and $\lambda_{\min}^+(\hat{\Sigma}_n)$ denote the largest and smallest nonzero eigenvalues, respectively, of $\hat{\Sigma}$, and the second inequality holds almost surely for large enough n, by the Bai-Yin theorem. Observe that

$$f_n(z) \to z S_\gamma(-z) r^2 \quad \text{pointwisely almost surely}$$

As its derivatives are bounded, the sequence f_n is equicontinuous, and by the Arzela-Ascoli theorem, we deduce that f_n converges uniformly to its limit. Since one of the limits converge uniformly we can exchange limits in (2) to get that

$$B_X(\hat{\beta}; \beta) \to r^2 \lim_{z \to 0^+} z S_\gamma(-z).$$

Then taking limits in (1) gives

$$\begin{aligned} \lim_{z \to 0^+} z S_\gamma(-z) &= \lim_{z \to 0^+} \frac{-(1 - \gamma + z) + \sqrt{(1 - \gamma + z)^2 + 4\gamma z}}{2\gamma} \\ &= \frac{-(1 - \gamma) + (\gamma - 1)}{2\gamma} = 1 - 1/\gamma \end{aligned}$$

□

3. Appendix

We now give a heuristic sketch for the proof of Theorem 7.6.

Proof of Rubio-Mestre. Without loss of generality suppose that $r^2 = 1$. Then by the definition of the resolvent $R_W(W - z) = I$ which implies that

$$\langle \beta, R_W W \beta \rangle - z \langle \beta, R_W \beta \rangle = 1$$

Taking expectations and use Sherman-Morrison,

$$\begin{aligned} z\mathbb{E}\langle \beta, R_W \beta \rangle + 1 &= \mathbb{E} \frac{\langle \beta, R_W X_1 X_1^T \beta \rangle}{1 + \frac{1}{M} \langle X_1 R_W X_1 \rangle} \\ &= \mathbb{E} \frac{\langle X_1, \beta \beta^T R_W X_1 \rangle}{1 + \frac{1}{M} \langle X_1, R_W X_1 \rangle}, \end{aligned}$$

by Lemma 3.4 from the previous lecture,

$$\begin{aligned} \langle X_1, \beta \beta^T R_W X_1 \rangle &\approx \operatorname{tr}(\beta \beta^T R_W \rangle \\ &= \langle \beta, R_W \beta \rangle \end{aligned}$$

Writing $\tilde{S} = \lim \mathbb{E}\langle \beta, R_W \beta \rangle = \lim \langle \beta, R_W \beta \rangle$ we have that

$$\begin{aligned} 1 &= \frac{\tilde{S}}{1 + \frac{1}{\gamma} \tilde{S}} - z\tilde{S} \\ &= \tilde{S} \left(\frac{1}{1 + \frac{1}{\gamma} S} - z \right) = \frac{\tilde{S}}{S} \end{aligned}$$

which implies that $S = \tilde{S}$ for which the result follows. □

CHAPTER 8

Neural Tangent Kernel

1. Overview: Training Dynamics

The remainder of this course will focus on understanding the training dynamics of neural networks. Given a loss function L and training dataset $\{(y_l, x_l)\}_{l=1}^n$, training a neural network often involves solving the random optimization problem

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{l=1}^{n} L(y_\ell, x_\ell; \theta)$$

where θ represents the network parameters, Θ is the space of possible parameter configurations we are optimizing over, and $\hat{R}(\theta) = \sum_{l=1}^n L(y_\ell, x_\ell; \theta)/n$ is the empirical risk. Note that the empirical risk is random because it is a function of the training dataset.

In contrast to classical optimization, we have additional challenges to consider when training neural networks:

(1) **Non-Convexity:** The loss landscapes for neural networks are notoriously non-convex. Even if we chose a convex loss function L, the mapping $\theta \mapsto \hat{R}(\theta)$ is non-convex due to the nonlinear activations introduced within the neural network. In particular, the spectrum of $\nabla_\theta^2 \hat{R}(\theta)$ (i.e. Hessian of the empirical risk) can have a number of negative eigenvalues on the order of the input dimension.
(2) **High-Dimensional:** The dimension of Θ, the optimization space, is of the order of millions or billions of parameters.

Despite this, first-order gradient descent algorithms are able to find solutions that both minimize the empirical risk and exhibit good generalization performance. Our goal then is to try and understand the trajectory of the training process. In particular, we will model the training dynamics using *gradient flow*:

$$\frac{d\theta_t}{dt} = -\nabla_\theta \hat{R}(\theta_t), \quad \theta_0 \sim \mu \in \mu_1(\Theta)$$

where $\mu_1(\Theta)$ is the collection of probability distributions defined over Θ and θ_0 is initialized randomly. This formulation allows us to analyze the training process as a high-dimensional *random dynamical system*. The question now is how can we analyze θ_t?

1.1. Approaches and Regimes.

There are numerous techniques that have been proposed to analyze θ_t, each dealing with different regimes and assumptions. Some notable examples are:

(1) **Neural Tangent Kernel (NTK):** This approach gained significant attention early on as it connects the behavior of wide neural networks to kernel methods. This will be the topic of this lecture.

(2) **Infinite-Width/Mean Field Theory/Hydrodynamics:** This line of work draws inspiration from techniques used to study particles in hydrodynamic systems.
(3) **High-Dimensional Limit Theory / Effective Dynamics / Dynamical Mean Field Theory:** These approaches look to understand training dynamics in the high-dimensional regime, where both the data dimension and the number of parameters are arbitrarily large. They often involve deriving "effective" equations that capture the essential features of the training dynamics.
(4) **Sample Complexity Bounds:** This perspective focuses on understanding the relationship between the number of training samples and the generalization performance of the trained model. For example, it aims to determine the minimum number of samples required to achieve a certain level of accuracy or generalization error.

The first three approaches are examples of **scaling limits**, where the idea is to study the behavior of the training process as certain variables tend to infinity. Doing so often yields a tractable description of the training dynamics which are amenable to rigorous analysis.

2. Lazy Training

We will take a closer look at the **lazy training regime** this week. This scaling regime corresponds to the scenario where we take the parameter dimension to infinity while keeping the other variables fixed. The reasoning is that we are fitting a model with sufficient capacity to interpolate the training data without its parameters shifting considerably from their initialization.

EXAMPLE 8.1. Consider training a neural network $f(\boldsymbol{x}_i;\theta)$ on a regression problem with squared ℓ_2 loss,

$$\hat{R}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i;\theta))^2$$

where $\theta \in \mathbb{R}^p$ is the parameter vector and $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \mathbb{R}$ is the training set. For notational brevity, we write the vector of predictions $f(\boldsymbol{x}_i;\theta)$ for all $i = 1,\dots,n$ as

$$\boldsymbol{f}_n = (f(\boldsymbol{x}_1;\theta),\dots,f(\boldsymbol{x}_n;\theta))$$

The gradient flow dynamics can be rewritten as:

$$\frac{d\theta_t}{dt} = -\nabla_\theta \hat{R}(\theta_t) = -\frac{1}{n}\mathrm{D}\boldsymbol{f}_n(\theta_t)^\top(\boldsymbol{y} - \boldsymbol{f}_n(\theta_t))$$

where $\boldsymbol{y} = (y_1,\dots,y_n)^\top$ is similarly defined as the vector of target values.

As a reminder, taking $p \to \infty$ implies we fit an interpolator with so many parameters that gradient flow can find a solution within some neighbourhood of the initialization $\theta_{\text{init}} = \theta_0$. If this is the case, we can *linearize* the dynamics and instead study the local behaviour of gradient flow about θ_0. Thus the *linearized dynamics* can be expressed as:

$$\frac{d\tilde{\theta}_t}{dt} = -\mathrm{D}\boldsymbol{f}_n(\theta_0)^T(\boldsymbol{y} - \boldsymbol{f}_n(\theta_0) - \mathrm{D}\boldsymbol{f}_n(\theta_0)(\tilde{\theta}_t - \theta_0))$$

where $\tilde{\theta}_t$ represents the linearized trajectory. Note that this corresponds to gradient flow on the "linearized" risk

$$\hat{R}^{\text{lin}}(\tilde{\theta}) = \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{f}_n(\theta_0) - \mathrm{D}\boldsymbol{f}_n(\theta_0)^\top(\tilde{\theta} - \theta_0)\|_2^2$$

which further corresponds to linearizing the original network

$$\boldsymbol{f}_n(\tilde{\theta}) \approx \boldsymbol{f}_n^{\text{lin}}(\tilde{\theta}) = \boldsymbol{f}_n(\theta_0) + \mathrm{D}\boldsymbol{f}_n(\theta_t)^\top(\theta_t - \theta_0)$$

Thus, under certain scaling regimes, we have that the linearized network is a good approximation of the original network, i.e. $\boldsymbol{f}_n \simeq \boldsymbol{f}_n^{\text{lin}}$. Furthermore, it has been established that under such scaling there is a connection between wide neural networks and kernel methods for which there is a larger body of established work to leverage insights from.

3. Kernel Methods

We will briefly review kernel methods and kernel gradient descent in this section before proceeding to the crux of the lecture on the neural tangent kernel.

Definition 8.2. A *kernel* is a positive semi-definite function $K(x, x')$ that measures a notion of similarity between two inputs x and x'.

Each kernel can be represented by its eigenbasis expansion as:

$$K(x, x') = \sum_i \lambda_i \phi_i(x)\phi_i(x')$$

where λ_i are the eigenvalues and $\phi_i(x)$ are the corresponding eigenfunctions (or *feature maps* in machine learning literature). Notably, these eigenfunctions $\{\phi_i\}$ span a special type of Hilbert space, called the **Reproducing Kernel Hilbert Space** (RKHS), where the inner product is induced by the kernel.

The goal of kernel methods is to approximate a target function $f(x)$ as a linear combination of these feature maps:

$$f(x) \approx \sum_i c_i \phi_i(x)$$

where the coefficients c_i are learned from the training data. To learn the parameters we can make use of a special form of gradient descent called *kernel gradient descent*. We will illustrate this using the previous example.

Example 8.3. Suppose we wish to learn the truncated kernel representation,

$$f(x; \theta) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \theta_i \phi_i(x)$$

Recall that for the least squares objective, the gradient flow dynamics are governed by:

$$\frac{d\theta_t}{dt} = -\nabla_\theta \hat{R}(\theta_t) = \frac{1}{2}\sum_\ell (y_\ell - f(\boldsymbol{x}_l; \theta))\nabla_\theta f(\boldsymbol{x}_l; \theta)$$

Now, rather than directly analyzing the evolution of θ_t, we can instead look at the evolution of the function output $f(x;\theta_t)$ at some new data point x. By the chain rule,

$$\frac{d}{dt}f(\boldsymbol{x};\theta_t) = \langle \nabla_\theta f(\boldsymbol{x};\theta_t), \dot{\theta}_t \rangle = -\frac{1}{2}\sum_{l=1}^{n}(y_l - f(x_l;\theta_t))\langle \nabla_\theta f(x_l;\theta_t), \nabla_\theta f(x;\theta_t)\rangle$$

However, by substituting the kernel representation of the gradient we obtain a new kernel that is independent of θ, i.e.

$$\langle \nabla_\theta f(x_l;\theta_t), \nabla_\theta f(x;\theta_t)\rangle = \frac{1}{m}\sum_{i=1}^{m}\phi_i(x_l)\phi_i(x) = K(x_l, x)$$

This leads to a simplified description of the dynamics as a function of the kernel:

$$\frac{d}{dt}f(\boldsymbol{x};\theta_t) = -\frac{1}{2}\langle K(\boldsymbol{x},\cdot), \boldsymbol{y} - \boldsymbol{f}_n(\boldsymbol{x};\theta_t)\rangle$$

which describes the evolution of the function output as a linear system involving the newly constructed kernel matrix.

3.1. Random Feature Regression. Rather than explicitly computing the kernel matrix, we can approximate a kernel function using *random feature maps*. In particular, we can construct such a map as follows:

- Fix $\sigma : \mathbb{R} \to \mathbb{R}$ and generate n random vectors $\{w_\ell\}_{\ell=1}^{n}$
- We define the ℓ-th feature in the map as $\phi_\ell(x) = \sigma(\langle w_\ell, x\rangle)$
- The (random) kernel induced by the random feature map ϕ is then given by

$$K_n(x, x') = \frac{1}{m}\sum_{\ell=1}^{m}\phi_\ell(x)\phi_\ell(x') = \frac{1}{m}\sum_{\ell=1}^{m}\sigma(\langle w_\ell, x\rangle)\sigma(\langle w_\ell, x'\rangle)$$

By the law of large numbers, we have that as $m \to \infty$,

$$K_n(x, x') \xrightarrow{P} \mathbb{E}_w[\sigma(\langle w, x\rangle)\sigma(\langle w, x'\rangle)] = K(x, x')$$

Random feature regression provides a link between kernel methods and neural networks. In particular for a sufficiently wide neural network at initialization, the first layer can be interpreted as a random feature map of the input for which subsequent layers learn a function of these features.

4. Neural Tangent Kernel

Consider an L-layer fully connected neural network with the following specifications:

- The number of neurons (i.e. width) of each layer is $n_0, n_1, \ldots, n_L$ where n_0 is the fixed input dimension and n_L is the output dimension.
- The non-linear activation $\sigma : \mathbb{R} \to \mathbb{R}$ is Lipschitz.
- The full network $f(x;\theta) = \tilde{\alpha}^{(L)}(x,\theta)$ is recursively defined as

$$\begin{aligned}
\alpha^{(0)}(x,\theta) &= x, \\
\tilde{\alpha}_{k+1}(x,\theta) &= \frac{1}{\sqrt{n_k}}W^{(k)}\alpha^{(k)}(x,\theta) + \beta b^{(k)} \\
\alpha^{(k)}(x,\theta) &= \sigma(\tilde{\alpha}^{(k)}(x,\theta))
\end{aligned}$$

where $\theta = \{W^{(k)}, b^{(k)}\}_{k=0}^{L}$ is the collection of network parameters (we can of course absorb the bias term into the weight matrix).

- Initialize the parameters with $W_{ij}^{(k)}, b_l^{(k)} \sim \mathcal{N}(0,1)$ iid (i.e. the entries of all parameters are independent standard Normal variables).
- The total number of parameters in the network is $p = \sum_{l=0}^{L-1} n_{l+1}(n_l + 1)$.

THEOREM 8.4 (Proposition 1 of Jacot et al. [2018]). *As $n_1, \ldots, n_{L-1} \to \infty$ with parameter initialization at θ_0, we have that $\tilde{\alpha}^{(k)}(x, \theta_0)$ for $k = 1, \ldots, L-1$ converges weakly to a centered Gaussian process with covariance function $\Sigma^{(L)}(x, x')$ defined recursively by*

$$\Sigma^{(1)}(x, x') = \frac{1}{n_0} x^T x' + \beta^2$$
$$\Sigma^{(k+1)}(x, x') = \mathbb{E}_{f \sim \mathcal{N}(0, \Sigma^{(k)}}[\sigma(f(x))\sigma(f(x'))] + \beta^2$$

The expectation can be interpreted as jointly taking two Gaussian random variables $z(x), z(x')$ within the nonlinearity σ with distribution

$$(z(x), z(x')) \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma^{(k)}(x,x) & \Sigma^{(k)}(x,x') \\ \Sigma^{(k)}(x',x) & \Sigma^{(k)}(x',x') \end{bmatrix}\right)$$

PROOF. We will prove this by induction on the number of layers L.

- **Base Case ($L = 1$):** For a single-layer network, we have that

  $$\tilde{\alpha}_1(\theta; x) = \frac{1}{\sqrt{n_0}} W^{(0)} x + \beta b^{(0)}$$

  where each component is just a linear combination of independent Gaussian variables. It follows that the covariance is

  $$\frac{1}{n_0}\mathbb{E}[(W^{(0)}x)^\top (W^{(0)}x')] + \beta^2 = \frac{1}{n_0} x^\top x' + \beta^2 = \Sigma^{(1)}(x, x')$$

- **Inductive Step:** At layer $k+1$, we have

  $$\tilde{\alpha}^{(k+1)}(x, \theta) = \frac{1}{\sqrt{n_k}} W^{(k)} \tilde{\alpha}^{(k)}(x, \theta) + \beta b^{(k)}$$

  By conditioning on the previous layer α_k, $\tilde{\alpha}^{(k+1)}$ becomes a centered Gaussian process with (conditional) covariance

  $$\tilde{\Sigma}^{(k+1)}(x, x') = \frac{1}{n_k} \langle \tilde{\alpha}^{(k)}(x, \theta), \tilde{\alpha}^{(k)}(x', \theta) \rangle + \beta^2$$

  By the inductive hypothesis, $\tilde{\alpha}^{(k)}(x, \theta)$ converges weakly to a Gaussian process with covariance $\Sigma^{(k)}(x, x')$, thus by the law of large numbers (taking $n_k \to \infty$),

  $$\tilde{\Sigma}^{(k+1)}(x, x') \to \mathbb{E}_{f \sim \mathcal{N}(0, \Sigma^{(k)})}[\sigma(f(x))\sigma(f(x'))] + \beta^2 = \Sigma^{(k+1)}(x, x')$$

  Since the converged quantity is independent of $\alpha^{(k)}$, it follows that the unconditioned feature map $\tilde{\alpha}^{(k+1)}$ also converges to the same covariance.

□

Up to this point we have only looked at the behaviour at initialization. To understand the evolution of the network *during* training, we need to make use of the empirical neural tangent kernel (NTK).

DEFINITION 8.5. The *empirical neural tangent kernel (NTK)* is defined as:

$$K_p^{(L)}(x,x') = \frac{1}{\sqrt{p}} \sum_{l=1}^{p} \nabla_{\theta^{(l)}} \tilde{\alpha}^{(L)}(x,\theta) \otimes \nabla_{\theta^{(l)}} \tilde{\alpha}^{(L)}(x',\theta)$$

We can also obtain similar results where the empirical NTK converges to a deterministic limiting kernel both at initialization and also during training.

THEOREM 8.6 (Theorem 1 of Jacot et al. [2018]). *As $n_1, \ldots, n_{L-1} \to \infty$, the empirical NTK at initialization converges in probability to a deterministic kernel*

$$K_P^{(L)}(x,x') \to K_\infty(x,x') \otimes I_{n_L}$$

where the limiting kernel $K_\infty(x,x')$ is defined recursively by

$$\begin{aligned} K_\infty^{(1)}(x,x') &= \Sigma^{(1)}(x,x') \\ K_\infty^{(k+1)}(x,x') &= K_\infty^{(k)} \dot{\Sigma}^{(k+1)}(x,x') + \Sigma^{(k+1)}(x,x') \\ \dot{\Sigma}^{(k+1)}(x,x') &= \mathbb{E}_{f\sim\mathcal{N}(0,\Sigma^{(k)})}[\dot{\sigma}(f(x))\dot{\sigma}(f(x'))] \end{aligned}$$

THEOREM 8.7 (Theorem 2 of Jacot et al. [2018]). *For any $T > 0$ such that the integral $\int_0^T \|\boldsymbol{y} - \boldsymbol{f}_n(x;\theta_t)\|^2 \, dt$ stays stochastically bounded, as $n_1, \ldots, n_{L-1} \to \infty$, we have that uniformly for $t \in [0,T]$,*

$$K_p^{(L)}(t) \to K_\infty^{(L)} \otimes I_{n_L}$$

Crucially, we see that at initialization and during training, the limiting kernel is deterministic and is independent of θ. Furthermore, the time-dependence of the empirical NTK during training on θ_t has washed out.

A key ingredient for proving Theorem 2 is the following lemma,

LEMMA 8.8 (Lemma 1 of Jacot et al. [2018]). *Under the setting of Theorem 2, for any $\ell = 1, \ldots, L-1$, we have the convergence in probability*

$$\lim_{n_{L-1}\to\infty} \cdots \lim_{n_1\to\infty} \sup_{t\in[0,T]} \left\| \frac{1}{\sqrt{n_\ell}} W^{(\ell)}(t) - W^{(\ell)}(0) \right\|_{op} = 0$$

This lemma is a consequence of "lazy training" and formally states that the intermediate weights essentially act as fixed, random feature maps (up to the scale $1/\sqrt{n_\ell}$) during training. We will briefly provide a proof sketch of this lemma in the context of the least-squares example to conclude the lecture.

PROOF. The idea is to introduce a differential inequality on the rates of all the weights and activations over all the layers and show that this bound diminishes as the network width increases.

We introduce the following quantities that depend on the re-scaled weights and activations

$$w_t^{(k)} = \left\| \frac{1}{\sqrt{n_k}} W^{(k)}(t) \right\|_{\text{op}}, \quad a_t^{(k)} = \sqrt{\frac{1}{n_k} \sum_{\ell=1}^{n} \alpha^{(k)}(x_\ell,\theta)^2} = \left\| \frac{1}{\sqrt{n_k}} \alpha_t^{(k)} \right\|_{p^{\text{in}}}$$

Finally, we introduce quantities that monitor the differences between the weights and activations during training,

$$\tilde{a}^{(k)}(t) := \left\| \frac{1}{\sqrt{n_k}} \left(\tilde{\alpha}^{(k)}(t) - \tilde{\alpha}^{(k)}(0) \right) \right\|_{p^{in}}, \quad \tilde{w}^{(k)}(t) := \left\| \frac{1}{\sqrt{n_k}} \left(W^{(k)}(t) - W^{(k)}(0) \right) \right\|_{\text{op}}$$

Thus we define a combined measure of all these quantities as

$$A(t) = \sum_{k=1}^{L} \left(a^{(k)}(0) + c\tilde{a}^{(k)}(t) + w^{(k)}(0) + \tilde{w}^{(k)}(t) \right)$$

The remainder of the proof is to control $A(t)$ using a differential inequality, specifically we can obtain one of the form

$$\dot{A}(t) \leq \frac{\mathcal{P}(A(t))}{\sqrt{\min\{n_1, \ldots, n_{L-1}\}}} \left\| \boldsymbol{y} - \boldsymbol{f}_n \right\|_{p^{in}}$$

where $\mathcal{P}(A(t))$ is a polynomial function of $A(t)$. The final step is to observe that $A(0)$ is also stochastically bounded in the limit and $A(t) \to A(0)$ which completes the proof. $\square$

CHAPTER 9

Lazy Training

1. Introduction

Consider a neural network with parametric form $f(\boldsymbol{x};\boldsymbol{\theta}) \in \mathbb{R}$ where $\{(y_i, \mathbf{x}_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ is the training dataset and $\boldsymbol{\theta} \in \mathbb{R}^p$ are the network parameters. We wish to train this network using the square loss function:

$$\hat{R}_n(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - f\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)\right)^2 = \frac{1}{2n}\left\|\boldsymbol{y} - \boldsymbol{f}_n(\boldsymbol{\theta})\right\|_2^2$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$, $\boldsymbol{f}_n = \left(f\left(\boldsymbol{x}_1;\boldsymbol{\theta}\right), \ldots, f\left(\boldsymbol{x}_n;\boldsymbol{\theta}\right)\right)$ are the vector representations of the target values and network outputs evaluated at the n data points in the training dataset (this is established notation from the previous lecture). Starting with an initial parameter vector $\boldsymbol{\theta}_0$, we can minimize the empirical risk using gradient flow:

$$\dot{\boldsymbol{\theta}}_t = -\nabla \hat{R}_n(\boldsymbol{\theta}) = \frac{1}{n}\mathrm{D}\boldsymbol{f}_n\left(\boldsymbol{\theta}_t\right)^\top\left(\boldsymbol{y} - \boldsymbol{f}_n\left(\boldsymbol{\theta}_t\right)\right) \tag{1.1}$$

where $\mathrm{D}\boldsymbol{f}_n\left(\boldsymbol{\theta}_t\right) \in \mathbb{R}^{n\times p}$ denotes the Jacobian matrix of $\boldsymbol{f}_n$.

In the previous lecture, one of the main results of Jacot et al. [2018] showed that in the infinite-width limit, $\boldsymbol{\theta}$ doesn't deviate far from its initialization $\boldsymbol{\theta}_0$ during training. This suggests that one can work with a linearized version of the gradient flow system obtained by taking the first-order Taylor expansion of the right-hand side of (1.1) around $\boldsymbol{\theta}_0$:

$$\dot{\bar{\boldsymbol{\theta}}}_t = \frac{1}{n}\mathrm{D}\boldsymbol{f}_n\left(\boldsymbol{\theta}_0\right)^\top\left(\boldsymbol{y} - \boldsymbol{f}_n\left(\boldsymbol{\theta}_0\right) - \mathrm{D}\boldsymbol{f}_n\left(\boldsymbol{\theta}_0\right)\left(\bar{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0\right)\right)$$

This linearized gradient flow corresponds to minimizing the linearized empirical risk:

$$\hat{R}^{\mathrm{lin}}(\bar{\boldsymbol{\theta}}) = \frac{1}{2n}\left\|\boldsymbol{y} - \boldsymbol{f}_n\left(\boldsymbol{\theta}_0\right) - \mathrm{D}\boldsymbol{f}_n\left(\boldsymbol{\theta}_0\right)\left(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\right\|_2^2$$

which is obtained by taking the first-order Taylor expansion of $\boldsymbol{f}_n(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$. Notably, we can rewrite the linearized risk as

$$\hat{R}^{\mathrm{lin}}(\bar{\boldsymbol{\theta}}) = \frac{1}{2n}\left\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\bar{\boldsymbol{\theta}}\right\|_2^2$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{f}_n\left(\boldsymbol{\theta}_0\right) - \mathrm{D}\boldsymbol{f}_n\left(\boldsymbol{\theta}_0\right)\boldsymbol{\theta}_0$ and $\tilde{\boldsymbol{X}} = \mathrm{D}\boldsymbol{f}_n\left(\boldsymbol{\theta}_0\right)$. Specifically, if the Jacobian has full row rank, then the set of global minimizers $\mathcal{E} = \{\bar{\boldsymbol{\theta}} : \hat{R}^{\mathrm{lin}}(\bar{\boldsymbol{\theta}}) = 0\}$ forms an affine space of dimension $p - n$. Furthermore, gradient flow converges to the solution $\bar{\boldsymbol{\theta}}_\infty \in \mathcal{E}$ which minimizes the problem,

$$\min_{\bar{\boldsymbol{\theta}}} \left\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_2^2 \quad \text{subject to} \quad \mathrm{D}\boldsymbol{f}_n\left(\boldsymbol{\theta}_0\right)\left(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \boldsymbol{y} - \boldsymbol{f}_n\left(\boldsymbol{\theta}_0\right)$$

In the lazy training regime, we expect that $\boldsymbol{\theta}_t$ is well-approximated by $\bar{\boldsymbol{\theta}}_t$ for all t. Thus the question now is to quantify any deviations between the two.

Before continuing, we define some useful quantities. Denote the Lipschitz constant of the Jacobian matrix by

$$L_n = \operatorname{Lip}(\mathrm{D}\boldsymbol{f}_n) := \sup_{\boldsymbol{\theta}\neq\boldsymbol{\theta}'} \frac{\|\mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}) - \mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}')\|_{\mathrm{op}}}{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2}$$

and further set $\sigma_{\min} = \sigma_{\min}(\mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_0))$ and $\sigma_{\max} = \sigma_{\max}(\mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_0))$.

THEOREM 9.1 (Theorem 5.1 from Bartlett et al. [2021] following Oymak and Soltanolkotabi [2019], Chizat et al. [2019]). *Suppose the following condition (we denote this as Assumption 1) holds:*

$$L_n \|\boldsymbol{y} - \boldsymbol{f}_n(\boldsymbol{\theta}_0)\|_2 < \frac{1}{4}\sigma_{min}^2$$

Then, for all $t > 0$, the following hold:

(1) $\widehat{R}(\boldsymbol{\theta}_t) \leq \widehat{R}(\boldsymbol{\theta}_0)\, e^{-\lambda_0 t}$ *with* $\lambda_0 = \sigma_{\min}^2/(2n)$

(2) $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \lesssim \dfrac{1}{\sigma_{\min}} \|\boldsymbol{y} - \boldsymbol{f}_n(\boldsymbol{\theta}_0)\|_2$

(3) $\left\|\boldsymbol{\theta}_t - \overline{\boldsymbol{\theta}}_t\right\|_2 \lesssim \left\{ \dfrac{\sigma_{max}}{\sigma_{min}^2} \|\boldsymbol{y} - \boldsymbol{f}_n(\boldsymbol{\theta}_0)\|_2 + \dfrac{L_n}{\sigma_{min}^3} \|\boldsymbol{y} - \boldsymbol{f}_n(\boldsymbol{\theta}_0)\|_2^2 \right\} \wedge \dfrac{L_n \sigma_{\max}^2}{\sigma_{\min}^5} \|\boldsymbol{y} - \boldsymbol{f}_n(\boldsymbol{\theta}_0)\|_2^2$

As a concrete example of how to apply the theorem, let's consider a family of two-layer neural networks with the following architecture:

$$f(\boldsymbol{x};\boldsymbol{\theta}) := \frac{\alpha}{\sqrt{m}} \sum_{j=1}^{m} b_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle), \quad \boldsymbol{\theta} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m) \in \mathbb{R}^{dm}, \quad b_j = \begin{cases} 1 & j \leq m/2 \\ -1 & j > m/2 \end{cases}$$

where m is the number of hidden units, σ is the activation function, and α is a scaling factor. The total number of parameters is $p = md$.

We will assume (denote this as *Assumption 2*) that the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is differentiable with bounded first and second-order derivatives. Furthermore, for some constant k_0, the coefficients of the first $k \leq k_0$ Hermite polynomials of σ are non-zero, i.e.

$$\mathbb{E}[\sigma(z)h_k(z)] \neq 0, \quad k \leq k_0$$

With this formulation, we can determine bounds on each of the quantities observed in the previous theorem as follows.

THEOREM 9.2 (Lemma 5.3 applied to Theorem 5.1 of Bartlett et al. [2021]). *Under Assumption 2, suppose the training data set $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ are i.i.d. with $\boldsymbol{x}_i \sim \mathcal{N}(0, \mathbf{I}_d)$ and y_i are B^2-sub-Gaussian. Furthermore, we initialize $\boldsymbol{\theta}_0 = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m)$ i.i.d. according to $\boldsymbol{w}_i \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. Then, provided that $md \gtrsim n \log n$ and $n \lesssim d^{k_0}$, with probability at least $1 - Ce^{-cn}$,*

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{f}_n(\boldsymbol{\theta}_0)\|_2 &\lesssim (1+\alpha)\sqrt{n} \\ \sigma_{\min} &\gtrsim \alpha\sqrt{d} \\ \sigma_{\max} &\lesssim \alpha(\sqrt{n} + \sqrt{d}) \\ \operatorname{Lip}(D\boldsymbol{f}_n) &\lesssim \alpha\sqrt{\frac{d}{m}}(\sqrt{n} + \sqrt{d}) \end{aligned}$$

If $\frac{n^2}{md} \lesssim C(\alpha)$ *and letting* $\lambda_* \asymp d/n$*, we have that with probability at least* $1 - \exp(-\Omega(n))$*,*

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \lesssim \sqrt{\frac{n}{d}}$$
$$\left\|\boldsymbol{\theta}_t - \overline{\boldsymbol{\theta}}_t\right\|_2 \lesssim \left\{\sqrt{\frac{n}{d}} + \sqrt{\frac{n}{md}}\right\} \wedge \frac{n}{d^{3/2}}$$

We will now prove Theorem 9.1.

PROOF. Let $\boldsymbol{y}_t = \boldsymbol{f}_n(\boldsymbol{\theta}_t)$ denote the network predictions at time t. Taking the time derivative and substituting the gradient flow equation for $\dot{\boldsymbol{\theta}}_t$,

$$\dot{\boldsymbol{y}}_t = \mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_t)\dot{\boldsymbol{\theta}}_t = -\frac{1}{n}\mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_t)\mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_t)^\top(\boldsymbol{y}_t - \boldsymbol{y})$$

Define the empirical kernel matrix at time t as $\boldsymbol{K}_t := \mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_t)\mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_t)^\top$. Then, the above equation becomes:

$$\dot{\boldsymbol{y}}_t = -\frac{1}{n}\boldsymbol{K}_t(\boldsymbol{y}_t - \boldsymbol{y})$$

Taking the time derivative of $\|\boldsymbol{y}_t - \boldsymbol{y}\|_2^2$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{y}_t - \boldsymbol{y}\|_2^2 = -\frac{2}{n}\langle \boldsymbol{y}_t - \boldsymbol{y}, \boldsymbol{K}_t(\boldsymbol{y}_t - \boldsymbol{y})\rangle$$

Now, define the stopping time $\tau = \inf\{t : \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 > \frac{\sigma_{\min}}{2L_n}\}$ as the first time the parameters move a distance $\frac{\sigma_{\min}}{2L_n}$ away from their initialization. For all $t < \tau$, we have $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \le \frac{\sigma_{\min}}{2L_n}$ implying that the minimum eigenvalue of $\boldsymbol{K}_t$ is bounded below by $\left(\frac{\sigma_{\min}}{2}\right)^2$. Therefore, for $t < \tau$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{y}_t - \boldsymbol{y}\|_2^2 \le -\frac{\sigma_{\min}^2}{2n}\|\boldsymbol{y}_t - \boldsymbol{y}\|_2^2 = -\lambda_0\|\boldsymbol{y}_t - \boldsymbol{y}\|_2^2$$

where $\lambda_0 = \frac{\sigma_{\min}^2}{2n}$. By Grönwall's inequality, we have that for $t \le \tau$,

$$\|\boldsymbol{y}_t - \boldsymbol{y}\|_2^2 \le \|\boldsymbol{y}_0 - \boldsymbol{y}\|_2^2 e^{-\lambda_0 t}$$

This proves that the residuals decreases exponentially fast with the desired rate λ_0 for $t \le \tau$. It remains to show that $\tau = \infty$ using contradiction.

Assume $\tau < \infty$. For any $t \le \tau$, we can establish that $\sigma_{\min}(\mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_t)) \ge \frac{\sigma_{\min}}{2}$ by applying the Hoffman-Wielandt inequality,

$$|\sigma_{\min}(\mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_t)) - \sigma_{\min}(\mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_0))| \le \|\mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_t)) - \mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_0))\|_F \le L_n\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \le \frac{\sigma_{\min}}{2}$$

Using the gradient flow equations, we have that for any $t \le \tau$

$$\|\dot{\boldsymbol{\theta}}_t\|_2 = \frac{1}{n}\|\mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_t)^\top(\boldsymbol{y}_t - \boldsymbol{y})\|_2$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{y}_t - \boldsymbol{y}\|_2 = -\frac{1}{n}\frac{\|\mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_t)^\top(\boldsymbol{y}_t - \boldsymbol{y})\|_2^2}{\|\boldsymbol{y}_t - \boldsymbol{y}\|_2} \le -\frac{\sigma_{\min}}{2n}\|\mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_t)^\top(\boldsymbol{y}_t - \boldsymbol{y})\|_2$$

By combining these inequalities,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\|\boldsymbol{y}_t - \boldsymbol{y}\|_2 + \frac{\sigma_{\min}}{2}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2\right) \le \frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{y}_t - \boldsymbol{y}\|_2 + \frac{\sigma_{\min}}{2}\left\|\dot{\boldsymbol{\theta}}_t\right\|_2 \le 0$$

which implies that

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \leq \frac{2}{\sigma_{\min}} \|\boldsymbol{y} - \boldsymbol{y}_0\|_2 < \frac{\sigma_{\min}}{2L_n}$$

which is a contradiction at $t = \tau$. The second statement follows from this bound as well.

To briefly sketch the proof for (3), let $\overline{\boldsymbol{y}}_t = \boldsymbol{f}_n(\boldsymbol{\theta}_0) + \mathrm{D}\boldsymbol{f}_n(\boldsymbol{\theta}_0)\left(\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0\right)$ which yields $\dot{\overline{\boldsymbol{y}}}_t = -\frac{1}{n}\boldsymbol{K}_0(\overline{\boldsymbol{y}}_t - \boldsymbol{y})$. Define the difference $\boldsymbol{r}_t = \boldsymbol{y}_t - \overline{\boldsymbol{y}}_t$. We then have

$$\dot{\boldsymbol{r}}_t = -\frac{1}{n}\boldsymbol{K}_t\boldsymbol{r}_t - \frac{1}{n}(\boldsymbol{K}_t - \boldsymbol{K}_0)(\overline{\boldsymbol{y}}_t - \boldsymbol{y})$$

which will yield the differential inequality (after some work bounding each of the individual quantities)

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{r}_t\|_2 \leq -\frac{\lambda_0}{2}\|\boldsymbol{r}_t\|_2 + \frac{C\sigma_{\max}L_n}{n}$$

The result follows from an extended exercise of Grönwall's inequality. □

The proof for Theorem 9.2 requires an understanding of the concentration of measure phenomenon.

2. Appendix: Theorems and Definitions

THEOREM 9.3 (Hoffman-Wielandt inequality). *If $A, B \in \mathcal{M}_{n,n}(\mathbb{C})$ are normal $n \times n$ matrices with respective eigenvalues $\lambda_1(A), \ldots, \lambda_n(A)$ and $\lambda_1(B), \ldots, \lambda_n(B)$, then, denoting by $\mathcal{S}_n$ the permutation group of $\{1, \ldots, n\}$,*

$$\min_{\pi \in \mathcal{S}_n} \sum_{i=1}^{n} \left|\lambda_i(A) - \lambda_{\pi(i)}(B)\right|^2 \leq \|A - B\|_F^2$$

THEOREM 9.4 (Grönwall's inequality). *Let $u : [0, T] \to \mathbb{R}$ be a real-valued differentiable function and β a real-valued continuous function defined over the same interval. If u satisfies the inequality*

$$u'(t) \leq \beta(t)u(t) \quad \text{for all } t \in (0, T)$$

Then

$$u(t) \leq u(0)\exp\left\{\int_0^t \beta(s)\,ds\right\} \text{ for all } t \in [0, T]$$

DEFINITION 9.5 (Sub-Gaussian Random Variables). A centered random variable Y is *B^2-sub-Gaussian* if $\mathbb{E}[e^{\lambda Y}] \leq e^{\lambda^2 B^2/2}$ for all $\lambda \in \mathbb{R}$.

CHAPTER 10

TBA

CHAPTER 11

Sample Complexity of SGD

1. Introduction

- **Parametric Family of Distributions**: Consider a parametric family of distributions $(P_x)_{x\in\mathbb{R}^N} \in \mathcal{M}_1(\mathbb{R}^{d_N})$, where $\mathcal{M}_1(\mathbb{R}^{d_N})$ denotes the set of probability measures on $\mathbb{R}^{d_N}$.
- **Data**: We observe $M = \alpha N$ i.i.d. samples $(Y^l)_{l=1}^M$ in $\mathbb{R}^{d_N}$, drawn from the distribution P_{θ_N}, referred to as the **data distribution**.
- **Objective**: Estimate the unknown parameter θ_N using these samples through online stochastic gradient descent (SGD).
- **Loss Function**: We have a loss function
$$L_N : \mathbb{R}^N \times \mathbb{R}^{d_N} \to \mathbb{R}.$$
- **SGD Update**: The online SGD with a constant learning rate δ_N is given by
$$X_l = X_{l-1} - \delta_N \nabla L_N(X_{l-1}, Y_l),$$
where the initialization X_0 is possibly random, with $X_0 \sim \mu_N \in \mathcal{M}_1(\mathbb{R}^N)$.

Goal: We aim to analyze the evolution of the sequence (X_l) in the high-dimensional regime as $N, d_N \to \infty$. This involves studying the behavior of the SGD iterates in high dimensions and examining the influence of the constant learning rate δ_N on the convergence of (X_l) towards an estimate of θ_N.

This type of high-dimensional asymptotic analysis is essential for understanding SGD behavior in large-scale problems, particularly under conditions where N and d_N grow at similar rates or according to the scaling factor α.

1.1. Heuristic Picture. The evolution of SGD can be heuristically divided into two phases:

- **Search Phase**: Initially, the algorithm "**wanders**" in a non-convex landscape, where the approximation to gradient descent on the population loss is relatively poor.
- **Descent Phase**: Later, SGD operates within a "**trust region**," descending quickly toward a local minimum. In this phase, it closely approximates gradient descent on the population loss, improving convergence to an optimal point.

Such a **burn-in** time perspective is implicit in approaches based on the ODE method (McLeish [1976]; Ljung [1977]). Precisely, consider the following linear interpolation
$$\tilde{X}_t^\delta = X_{\lfloor t/\delta \rfloor} + (t - \lfloor t/\delta \rfloor \delta) X_{\lfloor \frac{t}{\delta} \rfloor + 1}$$
Then $\tilde{X}_t^\delta \to \bar{X}_t$ weakly as $\delta \to 0$ which solves $\dot{\bar{X}}_t = -\nabla\phi(\bar{X}_t)$ with initialization $\bar{X}_0 = x_0$, where $\Phi(x) = \mathbb{E}(L(x, Y))$ is called the population loss that we will discuss further.

1.2. High-Dimensional Picture. In high-dimensional settings, the impact of the burn-in period becomes less negligible, as its dependence on dimensionality remains unclear. This observation motivates the following questions for our approach:

(1) Can one prove sample complexity bounds for SGD that are Poly(N)?
(2) What fraction of data is used/time is spent in the search phase vs. the descent phase?
(3) How does these answers change as one varies the loss function, $L_N(x, Y)$?

For the sake of simplicity, consider a parametric family of distributions $(P_x)_{x\in\mathbb{S}^{N-1}}$ of $\mathbb{R}^{d_N}$-valued random variables, parameterized by the unit sphere $\mathbb{S}^{N-1}$.

- **Data**: We are given $M = \alpha N$ i.i.d. samples $(Y^l)_{l=1}^M$ from the distribution P_{θ_N}, the *data distribution*, where $\theta_N \in \mathbb{S}^{N-1}$.
- **Objective**: Estimate the unknown parameter θ_N using these samples via online stochastic gradient descent (SGD).
- **Loss Function**: We define a loss function

$$L_N : \mathbb{S}^{N-1} \times \mathbb{R}^{d_N} \to \mathbb{R}.$$

- **Population Loss**: The population loss takes the form

$$\Phi_N(x) := \mathbb{E}_{Y\sim P_{\theta_N}}[L_N(x, Y)] = \phi(m_N(x)),$$

  where $m_N(x) = \langle x, \theta_N\rangle$ denotes the **correlation** between x and θ_N, and $\phi : [-1, 1] \to \mathbb{R}$.
- **SGD Initialization**: The initialization is $X_0 \sim \text{Unif}(\mathbb{S}^{N-1})$, where X_0 is drawn uniformly from the unit sphere $\mathbb{S}^{N-1}$.

Goal: We aim to analyze the evolution of the SGD sequence (X_l) in the high-dimensional limit as $N, d_N \to \infty$. This includes understanding how the high-dimensional setting impacts the behavior of the SGD iterates and the convergence towards θ_N.

2. Applications to Some Inference Problems

2.1. Online PCA. From a statistical perspective, a classical application of online SGD to a matrix model is online **Principal Component Analysis (PCA)**. Suppose that $(Y_l)_{l=1}^M$ are i.i.d. centered random vectors in $\mathbb{R}^N$ with covariance

$$\mathbb{E}[YY^T] = I + \lambda v_0 v_0^T,$$

where v_0 is a fixed unit vector ($\|v_0\|_2 = 1$), $\lambda > 0$ is a fixed signal-to-noise ratio. The objective is to iteratively optimize the loss

$$L(x, Y) = -\langle x, YY^T x\rangle$$

over $\mathbb{S}^{N-1}$. This can be viewed as a stochastic approximation to $-\langle x, \mathbb{E}[YY^T]x\rangle$.

We observe that

$$\Phi(x) = -\langle x, \mathbb{E}[YY^T]x\rangle = -\lambda\langle x, v_0\rangle^2 - 1.$$

2.2. Binary GMM. Consider maximum likelihood estimation for a two-component mixture of Gaussian distributions, assuming identical variances and known mixture weights. In the spherical case, the clusters are antipodal:

$$Y \sim p\mathcal{N}(v_0, I_d) + (1-p)\mathcal{N}(-v_0, I_d),$$

with $v_0 \in \mathbb{R}^N$ and $\|v_0\| = 1$. Without loss of generality, we set $v_0 = e_1$.

The log-likelihood is given by

$$\tilde{f}(x, Y) = \log\left(p\exp\left(-\frac{1}{2}\|Y-x\|^2\right) + (1-p)\exp\left(-\frac{1}{2}\|Y+x\|^2\right)\right).$$

Reparameterizing the mixture weight as $p = \frac{e^h}{e^h+e^{-h}}$ for some $h \in \mathbb{R}$, we find that maximum likelihood estimation corresponds to minimizing the loss

$$L(x, Y) = -\log\cosh(\langle Y, x\rangle + h).$$

By utilizing rotational invariance, the expected loss can be expressed as

$$\Phi(x) = -\mathbb{E}[\log\cosh(z_1 + \epsilon m + h)],$$

where $\epsilon = \pm 1$ with probabilities p and $1-p$, respectively.

2.3. Teacher-Student Network. Suppose that we are given some activation function $f : \mathbb{R} \to \mathbb{R}$, some fitting model $g : \mathbb{R} \to \mathbb{R}$, some feature vectors $(a^l)_{l=1}^M$, and with these, M noisy responses of the form

$$y^l = f(\langle a^l, v_0\rangle) + \epsilon^l$$

where $(\epsilon^l)_l$ are additive errors. Our goal is to estimate v_0 given this data. Our approach is by minimizing the least squares error. Consider as data the pairs $Y^l = (y^l, a^l)$ for $l = 1, \ldots, M$. Using student's model, $\hat{y}(x) = g(\langle a, x\rangle)$, one may write a quadratic loss L of the form,

$$L(x, Y) = L(x, (y, a)) = (y - g(\langle a, x\rangle))^2$$

Note that we may write the population loss as

$$\Phi(x) = \mathbb{E}\left[(f(\langle a, x\rangle) - g(\langle a, x\rangle))^2\right] + \mathbb{E}[\epsilon^2]$$

As we mentioned above, let us focus on the most studied regime, namely where (a^l) are i.i.d. standard Gaussian vectors in $\mathbb{R}^N$; for the (ϵ^l) we only assume they are i.i.d. mean zero with finite $4+\delta$-th moment for some $\delta > 0$.

Recall that the **Hermite polynomials**, which we denote by $(h_k(x))_{k=0}^\infty$, are the normalized orthogonal polynomials of the Gaussian distribution $\gamma(x) \propto \exp(-x^2/2)dx$. Define the k-th Hermite coefficient for an activation function $f \in L^2(\gamma)$ by

$$u_k(f) = \langle f, h_k\rangle_{L^2(\gamma)} = \frac{1}{\sqrt{2pi}}\int_{-\infty}^{\infty} f(z)h_k(z)e^{-z^2/2}dz$$

As long as f' is at most polynomial growth, the population loss is differentiable and the above exists. The population loss then has the following exact form

$$\Phi(x) = \phi_{f,g}(m(x))$$

where

$$\phi_{f,g}(m) := -2\sum_{k=0}^{\infty} u_k(f)u_k(g)m^k + \|f\|^2_{L^2(\gamma)} + \|g\|^2_{L^2(\gamma)} + C_\epsilon$$

Many commonly used activation functions are as follows,

- Well-specified: $f(x) = g(x)$
- Online linear regression: $f(x) = x$
- Phase retrieval: $f(x) = x^2$

In the well-specified case, one may simplify the above population loss. Let $v_0 \in \mathbb{S}^{N-1}$ be a fixed unit vector and the population loss is given by

$$\Phi(x) = \mathbb{E}\left[(f(\langle a, x\rangle) - f(\langle a, v_0\rangle))^2\right] + \mathbb{E}[\epsilon^2]$$

Then with the same conditions mentioned above,

$$\Phi(x) = \phi_f(m) := 2\sum_{k=0}^{\infty} u_k(f)^2(1 - m^k) + C_\epsilon$$

3. Preliminaries

3.1. Algorithm. As our parameter space is spherical (recall $\mathbb{S}^{N-1}$), we will consider a spherical version of online SGD which is defined as follows.

Definition 11.1 (Spherical Online SGD). Let X_t denote the output of the algorithm at time t, and let $\delta > 0$ denote a step size parameter. The sequence of outputs of the algorithm are then given by the following procedure:

$$\begin{cases} X_0 = x_0, \\ \tilde{X}_t = X_{t-1} - \frac{\delta}{N}\nabla L_N(X_{t-1}, Y^t) \\ X_t = \frac{\tilde{X}_t}{\|\tilde{X}_t\|} \end{cases}$$

where the initial point x_0 is possibly random, $x_0 \sim \mu \in \mathcal{M}_1(\mathbb{S}^{N-1})$ and where ∇ denotes the spherical gradient.

Remark 11.2. For a function $f : \mathbb{S}^{N-1} \to \mathbb{R}$, the spherical gradient is given by

$$\nabla f = Df - \frac{\partial f}{\partial r}\frac{\partial}{\partial r}$$

where D is the derivative and $\frac{\partial}{\partial r}$ is the partial derivative in the radial direction, both in $\mathbb{R}^N$.

We take our estimator, the output of this algorithm. In order for this algorithm to be well-defined, we assume that the loss is almost surely differentiable in the parameter for all $x \in \mathbb{S}^{N-1}$.

3.2. Weak and Strong Recovery. One of our goals is to understand the dimension dependence of the relative proportion of time spent by SGD in the search phase as opposed to the descent phase. To achieve this, we need a formal definition of the search phase:

- We focus on a simple setting where the population loss is a (possibly non-linear) function $f : [0, \infty) \to \mathbb{R}$ of the distance $d : \mathbb{S}^{N-1} \times \mathbb{S}^{N-1} \to [0, \infty)$ to the parameter, specifically defined as $\Phi(x) = f(d(x, x_0))$.
- Recall that the population loss can be expressed as $\Phi_N(x) = \phi(m_N(x))$.
- The correction $m_N(x)$ is often referred to as the **latitude**, and the set where $m_N(x) \approx 0$ is called the **equator** of the sphere.

- The reasoning behind this terminology is that with the random initialization $X_0 \sim \mu_N = \text{Unif}(\mathbb{S}^{N-1})$, we have $m_N(x_0) \approx \frac{1}{\sqrt{N}}$. By the **Poincaré limit**, it follows that $\sqrt{N} m_N(x_0) \to \mathcal{N}(0,1)$ as $N \to \infty$.

In order to formalize the notion of exiting the search phase, we recall here the notion of weak recovery, i.e., achieving macroscopic correlation with θ_N.

DEFINITION 11.3 (Weak Recovery). A sequence of estimators $\hat{\theta}_N \in \mathbb{S}^{N-1}$ is said to weakly recover (escape equator) the parameter θ_N if, for some $\eta > 0$, it holds that

$$\lim_{N\to\infty} P\left(m_N(\hat{\theta}_N) \geq \eta\right) = 1.$$

To understand this scaling here, i.e., $m_N(\hat{\theta}_N) = o(1)$, recall the basic fact from high-dimensional probability (Vershynin [2018]) that if $\hat{\theta}_N$ were drawn uniformly at random choice is exponentially small in N. In this context here, attaining weak recovery corresponds to exiting the search phase.

DEFINITION 11.4 (Strong Recovery). A sequence of estimators $\hat{\theta}_N \in \mathbb{S}^{N-1}$ is said to strongly recover (consistent estimaton) the parameter θ_N if it holds that

$$m_N(\hat{\theta}_N) \to 1, \quad \text{in probability (or equivalently in } L^p\text{-norm for every } p \geq 1).$$

In this context here, attaining weak recovery corresponds to the descent phase.

3.3. Assumptions. To investigate the performance of SGD in high-dimensional inference tasks, it is essential to choose an appropriate scaling for the fluctuations of the loss. These fluctuations are represented by the sample-wise error, defined as

$$H_N^l(x) := L_N(x, Y^l) - \Phi_N(x).$$

We aim for a scaling regime that avoids the issues outlined in the introduction. Consequently, we operate under an assumption that holds for the loss functions of many high-dimensional problems.

DEFINITION 11.5 (Regularization - Ben Arous, Gheissari, and Jagannath [2021]). Assumption A holds if there exists $C, \epsilon > 0$ such that the following two moment bounds hold for all N,

(1) $\sup_{x\in\mathbb{S}^{N-1}} \mathbb{E}[(\langle \nabla H_N(x), \theta_N\rangle)^2] \leq C$
(2) $\sup_{x\in\mathbb{S}^{N-1}} \mathbb{E}[\|\nabla H_N(x)\|^{4+\epsilon}] \leq C N^{\frac{4+\epsilon}{2}}$

REMARK 11.6. The scaling relation between (1) and (2) is tight when ∇H_N is an i.i.d. sub-Gaussian vector. i.e.,

- $\mathbb{E}[(\langle \nabla H_N(x), \theta_N\rangle)^2] \leq C$ for some $C > 0$ and all $\theta_N \in \mathbb{S}^{N-1}$.
- $\mathbb{E}\left[\|\nabla H_N(x)\|^{4+\epsilon}\right] \leq \sqrt{\mathbb{E}\left[\sum_i (\partial_i H_N(x))^2\right]^{\frac{4+\epsilon}{2}}} = O\left(N^{\frac{4+\epsilon}{2}}\right)$ by Jensen's inequality.

As we examine a first-order method, it is reasonable to expect that the output of gradient flow serves as a **Fisher consistent** estimator for all random initial data with positive correlation, implying consistency when assessed on the population loss. Additionally, if ϕ is an even function, then $\phi'(0) = 0$. Accordingly, we introduce the following assumption.

DEFINITION 11.7 (Fisher consistent - Ben Arous, Gheissari, and Jagannath [2021]). Assumption B holds if ϕ is differentiable and $\phi'(t) < 0$ for all $t \in (0, 1)$.

REMARK 11.8. Assumption A holds if and only if gradient flow for the population loss eventually produces a consistent estimator, when started anywhere on the upper half-sphere $\{x : m_N(x) > 0\}$.

4. Information Exponent

In the following we show that a key quantity governing the performance of online SGD is the notion, which we call the **information exponent** for a population loss.

DEFINITION 11.9 (Information Exponent). We say that a population loss Φ_N has information exponent k if $\phi \in C^{k+1}([-1, 1])$ and there exist $C, c > 0$ such that

$$\begin{cases} \frac{d^l \phi}{dm^l}(0) = 0 & 1 \le l < k \\ \frac{d^k \phi}{dm^k}(0) \le -c < 0 \\ \| \frac{d^{k+1} \phi}{dm^{k+1}}(m) \|_\infty \le C \end{cases}$$

The first theorem establishes an upper bound on the sample complexity for consistent estimation. Specifically, we have:

- We define
$$\alpha_c(N, k) = \begin{cases} 1 & k = 1 \\ \log N & k = 2 \\ N^{k-2} & k \ge 3 \end{cases}$$
- A sequence x_n is denoted as $x_n \ll y_n$ if $\frac{x_n}{y_n} \to 0$.
- For clarity, we present the following result under the case of random initialization, where μ_N is the uniform measure conditioned on the upper half sphere $\{m(x) \ge 0\}$. This conditioning is made without loss of generality up to a probability of $1/2$ and is implemented to mitigate symmetry issues.

THEOREM 11.10 (Ben Arous, Gheissari, and Jagannath [2021]). *Suppose that Assumptions A and B hold and that the population loss has information exponent k. Let $M = \alpha_N N$ with α_N growing at most polynomial in N, namely $\alpha_N = Poly(N)$. If (α_N, δ_N) satisfy $\alpha_N^{-1} \ll \delta_N \ll \alpha_N^{-1/2}$ and*

$$\alpha_N \gg \begin{cases} \alpha_c & k = 1 \\ \alpha_c \log N & k = 2 \\ \alpha_c (\log N)^2 & k \ge 3 \end{cases}$$

then online SGD with the learning rate δ_N started from $X_0 \sim \mu_N$, will have

$$m_N(X_M) \to 1, \quad \text{in probability and } L^p\text{-norm for every } p \ge 1.$$

The second theorem is the correspondent lower bound on the sample complexity required for exiting the search phase with a given information exponent.

THEOREM 11.11 (Ben Arous, Gheissari, and Jagannath [2021]). *Suppose that Assumptions A and B hold and that the population loss has information exponent k. If (α_N, δ_N) are such that*

- $(k \leq 2) : \alpha_N \ll \alpha_c$, *and* $\delta_N = O(1)$
- $(k \geq 3) : \alpha_N \ll \alpha_c$, *and* $\delta_N = O(\alpha_N^{-1/2})$,

then the online SGD with the learning rate δ_N, *started from* $X_0 \sim \mu_N$, *will have*

$$\sup_{t \leq M} |m_N(X_t)| \to 0, \quad \textit{in probability and } L^p\textit{-norm for every } p \geq 1.$$

The final argument demonstrates that the ratio of the number of samples utilized in the descent phase to those used in the search phase is $O(\alpha_c(N,k)^{-1})$, which approaches zero for $k \geq 2$. This observation is derived from Theorem 4.3 along with the following considerations.

THEOREM 11.12 (Ben Arous, Gheissari, and Jagannath [2021]). *Let* τ_η^+ *denote the first time* t *such that* $m_N(X_t) > \eta$ *(exit time), and* $\tau_{1-\eta}^+$ *the first time* t *such that* $m_N(X_t) > 1-\eta$. *Assume that Assumptions A and B hold and the population loss has information exponent* $k \geq 2$.

Define $M = \alpha_N N$ *with* (α_N, δ_N) *as in Theorem 4.2. For any* $\eta > 0$, *there exists a constant* $C = C(k,N) > 0$ *such that:*

$$\tau_\eta^+ \gg \alpha_c(N,k),$$

$$|\tau_{1-\eta}^+ - \tau_\eta^+| \leq CN \quad \textit{with probability } 1 - o(1).$$

Additionally, for all $\tau_{1-\eta}^+ \leq t \leq M$,

$$X_t > 1 - 2\eta \quad \textit{with probability } 1 - o(1).$$

In words Theorem 4.4 says that most of the data is used in the search phase (i.e. to attain some non-trivial correlation), and that from there descent to essentially full correlation is rapid, and takes $O(N)$ samples independently of the class of the problem.

5. Proof Idea

Here, we discuss the intuition behind the information exponent and Theorems 4.2 and 4.3. These theorems collectively demonstrate that the information exponent sharply dictates the performance of online SGD when initialized uniformly at random. (In fact, as shown in Theorems 4.2 and 4.3, this holds more generally for any initialization near the equator.)

When one considers the true online SGD, there is an effect due to the sample-wise error for the loss, H_N , whereby to first approximation,

$$m_t \approx m_{t-1} - \frac{\delta}{N}\langle \nabla \Phi, \nabla m \rangle - \frac{\delta}{N}\langle \nabla H_N, \nabla m \rangle$$

Now consider the case m is small, then

$$\langle \nabla \Phi, \nabla m \rangle = \phi'(m)\|\nabla m\|^2 \approx -a_k m^{k-1}$$

Due to the independence of ∇H_N^t and X_{t-1}, as we sum the contributions of the third term in time, we obtain a martingale which we call the **directional error martingale**,

$$M_t = \frac{\delta}{N}\sum_{j=0}^{t-1}\langle \nabla H_N^j(X_{j-1}), \theta_N \rangle$$

Therefore, one may write,

$$m_t = m_0 + a_k \frac{\delta}{N} \sum_{j=0}^{t-1} -M_t$$

Using Assumption A and Doob's inequality, its cumulative effect can be seen to typically be of the given order,

$$\sup_{t \leq T} |M_t| \lesssim \frac{\delta}{N} \sqrt{T}$$

In order for this term's contribution to be negligable for time scales on the order of M, and allow for recovery, we ask that

$$T = M = \alpha N \lesssim \frac{\delta}{N} \sqrt{\alpha} < c m_0$$

Hence,

$$m_t \gtrsim \frac{m_0}{2} + \frac{a_k \delta}{N} \sum_{j=0}^{t-1} m_j^{k-1}$$

Now we may use the discrete Bihari-LaSalle/Grownwall inequality: suppose that (m_t) is any sequence such that for some $a, b \geq 0$,

$$m_t \geq a + b \sum_{j=0}^{t-1} m_j^{k-1}$$

Applying the above inequality, one may write

$$m_t \geq \begin{cases} a(1+b)^t & k = 2 \quad \text{Gronwell's inequality} \\ a/(1 - ba^{k-2}t)^{1/(k-2)} & k \geq 3 \quad \text{Bihari-LaSalle inequality} \end{cases}$$

So if $a \asymp \frac{1}{\sqrt{N}}$ and $b \asymp \frac{\delta}{N}$, for the case of $k = 2$, we know that $m_t \gtrsim \frac{1}{\sqrt{N}} e^{t\delta/N}$ yields that we need $t \asymp N \log N$, and for the case of $k \geq 3$, we need of similar. However, note that Bihari-LaSalle inequality has finite time blow up but $|M| \leq 1$ so the above inequality only holds if $m \gtrsim \eta$ which needs differential analysis by Theorem 4.4

CHAPTER 12

Effective Dynamics

1. Introduction

- **Parametric Family of Distributions**: We have a family of distributions $(P_x)_{x\in\mathbb{R}^{p_N}} \subset \mathcal{M}_1(\mathbb{R}^{d_N})$, where $\mathcal{M}_1$ denotes the set of probability measures on $\mathbb{R}^{d_N}$.
- **Data**: A sequence of $M = \alpha N$ i.i.d. samples $(Y^l)_{l=1}^M$ in $\mathbb{R}^{D_N}$ is drawn from the distribution P_{θ_N}, called the **data distribution**.
- **Objective**: Estimate the unknown parameter θ_N using the samples via online stochastic gradient descent (SGD).
- **Loss Function**: The loss function is defined as

$$L_N : \mathbb{R}^{p_N} \times \mathbb{R}^{d_N} \to \mathbb{R}.$$

- **SGD Update**: The online SGD with a constant learning rate δ_N is given by:

$$X_l = X_{l-1} - \delta_N \nabla L_N(X_{l-1}, Y_l),$$

where X_0 is initialized, potentially at random, with $X_0 \sim \mu_N \in \mathcal{M}_1(\mathbb{R}^{p_N})$.
- **Goal**: To study the evolution of the sequence (X_l) as $p_N, D_N \to \infty$ when $N \to \infty$.

1.1. Classical Asymptotic Theory.

- **Classical Theory in Fixed Dimensions**: The asymptotic behavior of SGD and stochastic approximation has been well-established, with key results including: Pathwise limit theorems, such as functional central limit theorems and large deviation principles (Robbins and Monro [1951]; Ljung [1977]; McLeish [1976]).
- **Linear Interpolation**: For the linear interpolation

$$\tilde{X}_t^\delta = X_{\lfloor t/\delta \rfloor} + (t - \lfloor t/\delta \rfloor \delta) X_{\lfloor t/\delta \rfloor + 1},$$

we have that $\tilde{X}_t^\delta \to \bar{X}_t$ weakly as $\delta \to 0$.
- The limit $\bar{X}_t$ satisfies the differential equation

$$\dot{\bar{X}}_t = -\nabla \phi(\bar{X}_t),$$

with initial condition $\bar{X}_0 = x_0$.
- **Large Deviations**: Based on large deviation principles, the following probability inequality holds:

$$P(\|X_t^\delta - \dot{\bar{X}}_t\|_\infty > \epsilon) \approx e^{-I(\epsilon)/\delta}.$$

1.2. High-Dimensional Theory. In this lecture, we study a unified approach to the scaling limits of SGD in high dimensions with a constant learning rate, which allows us to understand a broad range of estimation tasks. More precisely, suppose $p, \delta \to \infty$ while $\delta \downarrow 0$; for example, $d = p^\alpha$ and $\frac{1}{\delta} = p^\beta$ for some $\alpha, \beta > 0$.

One, of course, cannot develop a high-dimensional scaling limit for the full trajectory of SGD as the dimension of the underlying parameter space is growing. However, in practice, one is rarely interested in the full trajectory. Instead, one typically tracks the evolution of certain weights or correlations between the classifier and the ground truth in a supervised setting.

We will see that, under mild regularity assumptions, the evolution of the **summary statistics** converges as the dimension grows to the solution of a system of (possibly stochastic) differential equations. These effective dynamics depend significantly on:

- The initialization of the parameters.
- The parameter regions in which one is developing the scaling limit.
- The scaling of the step-size with the dimension.

In practice, SGD often exhibits two types of phases in training:

- **Ballistic phases**, where the summary statistics change macroscopically, and
- **Diffusive phases**, where the summary statistics fluctuate microscopically.

We will study scaling limits for both types of phases.

2. Preliminaries

2.1. Summary Statistics. Suppose that we are given a sequence of functions $\mathbf{u}_N \in C^1(\mathbb{R}^{p_N}, \mathbb{R}^k)$ for some fixed k, where $\mathbf{u}_N(x) = (u_1^N(x), \dots, u_k^N(x))$. There are several important summary statistics, including norms, covariance, inner products, and others. Our goal is to understand the evolution of $\mathbf{u}_N(X_l)$ and build up a scaling limit for $\mathbf{u}_n(t) = \mathbf{u}_n(X_t^\delta) \to \mathbf{u}(t)$.

2.2. Localizability. To develop a scaling limit, we need some regularity assumptions on the relationship between how the step-size scales in relation to the loss, L, its gradients, ∇L, and the data distribution $\mathbb{P}$. To this end, we need to choose a scaling for the fluctuations of the loss. These fluctuations are captured by the sample-wise error, defined by

$$H(x, Y) := L_N(x, Y) - \Phi_N(x).$$

We suppress the dependence of H on Y and instead view H as a random function of x, denoting it $H(x)$. We let $G(x) = \mathbb{E}[\nabla H(x) \otimes \nabla H(x)]$ be the covariance matrix, also known as the **Fisher information** for ∇H at x.

We make two assumptions on the triple $(\mathbf{u}_N, L_N, P_{\theta_N})$ and the step size δ_N. The first is an upper bound on the learning rate in terms of regularity of the summary statistics. The second is the key assumption that the summary statistics' evolutions are asymptotically close. These assumptions need not hold uniformly over the entire parameter space $\mathbb{R}^{p_N}$, but only uniformly over the pre-image of compact sets under $\mathbf{u}_N$.

Definition 12.1 (Localizability). A triple $(\mathbf{u}_N, L_N, P_{\theta_N})$ is δ_N-localizable with localizing sequence $(E_K)_K$ of $\mathbb{R}^k$, and constants C_K (independent of N) such that

(1) $\max_i \sup_{x \in \mathbf{u}_n^{-1}(E_K)} \|\nabla^2 u_i^N\|_{op} \le C_K \delta_N^{-1/2}$

(2) $\begin{cases} \sup_{x \in \mathbf{u}_n^{-1}(E_K)} \|\nabla \Phi\| \le C_K \\ \sup_{x \in \mathbf{u}_n^{-1}(E_K)} \mathbb{E}[\|\nabla H\|^8] \le C_K \delta_N^{-4} \end{cases}$

$$(3)\quad \begin{cases} \max_i \sup_{x\in \mathbf{u}_n^{-1}(E_K)} \mathbb{E}[\langle \nabla H, \nabla u_i^N\rangle^4] \le C_K \delta_n^{-2} \\ \max_i \sup_{x\in \mathbf{u}_n^{-1}(E_K)} \mathbb{E}[\langle \nabla^2 H, \nabla H \otimes \nabla H - V\rangle^2] = o(\delta_N^{-3}) \end{cases}$$

REMARK 12.2. Suppose $\delta'_N \le \delta_N$. If δ_N is localizable, then δ'_N is localizable (Monotonicity).

2.3. The Martingale Problem Approach. In the following stochastic differential equation (SDE) we will see that there is a term, so called the drift term which can be seen from a Taylor expansion out to second order, with the role played by δ_N-localizability being to justify neglecting certain negligible second order terms, as well as all higher order terms. The identification of the stochastic term is via the classical martingale problem (Stroock and Varadhan [1997]) for summary statistics of SGD in the high-dimensional $N \to \infty$ limit. In other words, the convergence of Markov processes is equivalent with the convergence of elliptic operators.

To this end, we now turn to the second assumption, that the limiting evolution equations for the family of summary statistics chosen close. Define the following first and second-order differential operators,

$$\mathcal{A}_N = \sum_i \partial_i \Phi \partial_i, \quad \text{and} \quad \mathcal{L}_N = \frac{1}{2}\sum_{i,j} V_{ij}\partial_i\partial_j$$

Alternatively written $\mathcal{A}_N = \langle \nabla\Phi, \nabla\rangle$ and $\mathcal{L}_N = \frac{1}{2}\langle G, \nabla^2\rangle$. Moreover, we define the Jacobian operator as $J_n = \nabla \mathbf{u}_N$, which will be used with the Fisher information in the following sections.

DEFINITION 12.3 (Asymptotic Closability). A family of summary statistics $(\mathbf{u}_N)$ are asymptotically closable for learning rate δ_N if $(\mathbf{u}_N, L_N, P_{\theta_N})$ are δ_N-localizable with localizing sequence $(E_K)_K$, and furthermore there exist locally Lipschitz functions $\mathbf{h} : \mathbb{R}^k \to \mathbb{R}^k$ and $\mathbf{\Sigma} : \mathbb{R}^k \to \mathcal{P}_k \subseteq \mathbb{R}^k \times \mathbb{R}^k$, such that

$$\sup_{x\in \mathbf{u}_N^{-1}(E_K)} \|(-\mathcal{A}_N + \delta_N \mathcal{L}_N)\mathbf{u}_N(x) - \mathbf{h}(\mathbf{u}_N(x))\| \to 0$$

$$\sup_{x\in \mathbf{u}_N^{-1}(E_K)} \|\delta_N J_N G J_N^T - \mathbf{\Sigma}(\mathbf{u}_N(x))\| \to 0$$

In this case we call $\mathbf{h}$ the effective drift, and $\mathbf{\Sigma}$ the effective volatility.

Most of the assumed bounds in the definition of localizability are used to establish tightness and ensure higher order terms in Taylor expansions vanish in the $N \to \infty$ limit. In principle, the second assumption could not be dropped; in that case, the same quantity is still ensured to be $O(\delta^{-3})$ by the other localizability assumptions, but the limiting diffusion matrix would oppose simply the limit of $\delta J V J^T$.

3. Effective Dynamics

3.1. Main Result. We are now ready to present our main result which can be seen as a version of the classical martingale problem for high-dimensional SGD. For a function f and measure μ we let $f_*\mu$ denote the push-forward of μ.

THEOREM 12.4 (Ben Arous, Gheissari, and Jagannath [2024]). *Let $(X_l^{\delta_N})_l$ be SGD initialized from $X_0 \sim \mu_N$ for $\mu_N \in \mathcal{M}_1(\mathbb{R}^{p_N})$ with learning rate δ_N for the loss L_N and data distribution P_{θ_N}. For a family of summary statistics $\mathbf{u}_N = (u_i^N)_{i=1}^k$, let $(\mathbf{u}_N(t))_t$ be the linear interpolation of $(\mathbf{u}_N(X^{\delta_n}_{\lfloor t\delta_n^{-1}\rfloor}))_t$.*

Suppose that $\mathbf{u}_N$ are asymptotically closable with learning rate δ_N, effective drift $\mathbf{h}$, and effective volatility $\mathbf{\Sigma}$, and that the pushforward of the initial data has $(\mathbf{u}_N)_*\mu_N \to \nu$ weakly for some $\nu \in \mathcal{M}_1(\mathbb{R}^k)$. Then $(\mathbf{u}_N(t))_t \to (\mathbf{u}_t)_t$ weakly as $N \to \infty$ where $\mathbf{u}_t$ solves

$$d\mathbf{u}_t = \mathbf{h}(\mathbf{u}_t)dt + \sqrt{\mathbf{\Sigma}(\mathbf{u}_t)}d\mathbf{B}_t$$

initialized from ν, where $\mathbf{B}_t$ is a standard Brownian motion in $\mathbb{R}^k$.

We call the solution to the above SDE the **effective dynamics** of the summary statistics $\mathbf{u}_N$. The fact that $\mathbf{h}, \mathbf{\Sigma}$ are locally Lipschitz ensures that this solution is unique.

3.2. Comparison to Classical Theory. Let us compare the main result with the classical limit theory of stochastic gradient descent (SGD) in fixed dimension. For the sake of this discussion, suppose that each of the two terms $\mathcal{A}_N\mathbf{u}_N$ and $\delta_N\mathcal{L}_N\mathbf{u}_N$ individually admit $N \to \infty$ limits: namely, that there exist functions $\mathbf{f}, \mathbf{g} : \mathbb{R}^k \to \mathbb{R}^k$ such that

$$\sup_{x \in \mathbf{u}_N^{-1}(E_K)} \|\mathcal{A}_N\mathbf{u}_N(x) - \mathbf{f}(\mathbf{u}_N(x))\| \to 0, \tag{3.1}$$

$$\sup_{x \in \mathbf{u}_N^{-1}(E_K)} \|\delta_N\mathcal{L}_N\mathbf{u}_N(x) - \mathbf{g}(\mathbf{u}_N(x))\| \to 0. \tag{3.2}$$

In this case, the main result holds with $\mathbf{h} = -\mathbf{f} + \mathbf{g}$, and we call $\mathbf{f}, \mathbf{g}$, and $\mathbf{\Sigma}$ the **population drift**, the **population corrector**, and the **diffusion matrix** of $\mathbf{u}$ respectively.

From the fixed-dimensional perspective, when the first equation for $\mathbf{f}$ holds, one predicts $\mathbf{u}$ to solve

$$d\mathbf{u}_t = -\mathbf{f}(\mathbf{u}_t)dt$$

with initial data $\mathbf{u}_0 \sim \mathbf{u}_*\mu$, as this is its evolution under gradient descent on the population loss Φ.

Evidently, this perspective only applies in the high-dimensional limit (for $\delta_N' \ll \delta_N$) if both the population corrector $\mathbf{g}$ and the diffusion matrix $\mathbf{\Sigma}$ are zero.

- **Sub-Critical Scaling Regime:** For any triple $(\mathbf{u}_N, L_N, P_{\theta_N})$, there is a scaling of the learning rate δ_N with N below which $\mathbf{g} = \mathbf{\Sigma} = 0$, and the effective dynamics agree with the population loss. In this regime, the classical perspective applies.
- **Critical Scaling Regime:** In this case, $\mathbf{g}$ and $\mathbf{\Sigma}$ may be non-zero, indicating that high-dimensionality induces non-trivial corrections to $\mathbf{f}$. This situation raises an ongoing open problem regarding its relation to the "Edge of Stability".
- **Ballistic Limit:** The diffusion matrix for the effective dynamics of the most natural choice of summary statistics is zero ($\mathbf{\Sigma} = 0$) even in the critical scaling regime where $\mathbf{h} \neq \mathbf{f}$. In this case, the effective dynamics of the summary statistics is given by the ODE system:

$$d\mathbf{u}_t = \mathbf{h}(\mathbf{u}_t)dt.$$

- **Diffusive Limit:** If $\mathbf{\Sigma} > 0$, using the main result allows one to develop a separate diffusive limit.

4. Application to Some High-Dimensional Problems

4.1. Spiked Matrix Model. Consider the problem of de-noising a rank one matrix that has been corrupted additively by Gaussian noise via SGD. A popular statistical model of this task is the spiked matrix model of PCA. Suppose that we are given i.i.d. samples of data of the form

$$Y^l = \lambda v v^T + W^l$$

where W^l are i.i.d. copies of a k-matrix whose entries are i.i.d. standard Gaussians, $v \in \mathbb{R}^N$ is a unit vector, and $\lambda = \lambda_N > 0$ is the signal-to-noise ratio. Our goal is to infer v.

We take as loss the (negative) log-likelihood namely

$$L(x, Y) = \|Y - xx^T\|^2$$

The pair

$$m = m(x) := \langle x, v \rangle, \quad \text{and} \quad r_\perp^2 = r_\perp^2(x) := \|x\|^2 - m^2$$

are such that

$$\Phi(x) = \phi(m, r_\perp^2) = -2\lambda m^k + (r_\perp^2 + m^2)^k + c$$

and the law of L only depends on them.

In our normalization with $\lambda > 0$ fixed, the regime $\delta_N = o(1/N)$ is sub-critical and the regime $\delta_N = \Theta(1/N)$ is critical. We focus our presentation on the most interesting regime, namely the critical regime of $\delta_N = c_\delta/N$ for some constant c_δ. Note, however, that the limits in the subcritical regime are in all cases recovered by taking the $c_\delta \downarrow 0$ limits of the ODE/SDE's of the critical regime.

For notational simplicity, let $R^2 := m^2 + r_\perp^2$. We consider the pair $\mathbf{u}_N = (u_1, u_2) = (m, r_\perp^2)$, for which the main result yields the following effective dynamics (Ben Arous, Gheissari, and Jagannath [2024])

$$dm = 2m(\lambda k m^{k-2} - kR^{2(k-1)})dt, \quad dr_\perp^2 = -4kR^{2(k-1)}(r_\perp^2 - c_\delta)dt$$

where the above ODE initialized from $\lim_{N\to\infty}(\mathbf{u}_N)_*\mu_N$, and the c_δ term in the dynamic of $r_\perp^2$ is the population corrector.

If we focus on the critical step-size regime with $c_\delta = 1$ where one sees from the ODE's that $r_\perp^2 \to 1$ which is most directly related to an eigenvalue problem. The above effective dynamic for m and $r_\perp^2$ has isolated fixed points classified as follows:

(1) An unstable fixed point at $(0, 0)$
(2) A fixed point at $(1, 0)$ which is stable if $\lambda < 1$ and unstable if $\lambda > 1$.
(3) If $\lambda > 1$, a stable fixed point at $(1, 1)$.

4.2. BBP Transition. Let us now consider a rescaling of $\mathbf{u}_N$ in a microscopic neighborhood of the saddle set $m = 0$. This captures the initial phase from a random start: if $\mu_N \sim \mathcal{N}(0, I_N/N)$, then $(\mathbf{u}_N)_*\mu_N \to \delta_{(0,1)}$ weakly. Now rescale and let $\tilde{\mathbf{u}}_N = (\tilde{m}, r_\perp^2) = (\sqrt{N}m, r_\perp^2)$. Evidently, $\tilde{\nu} = \lim_{N\to\infty}(\tilde{\mathbf{u}}_N)_*\mu_N = \mathcal{N}(0, 1) \otimes \delta_1$. Then the effective dynamic (Ben Arous, Gheissari, and Jagannath [2024]) is given by

$$d\tilde{u}_1 = 4\tilde{u}_1(\lambda - \tilde{u}_2)dt + 2\sqrt{2\tilde{u}_2}dB_t$$

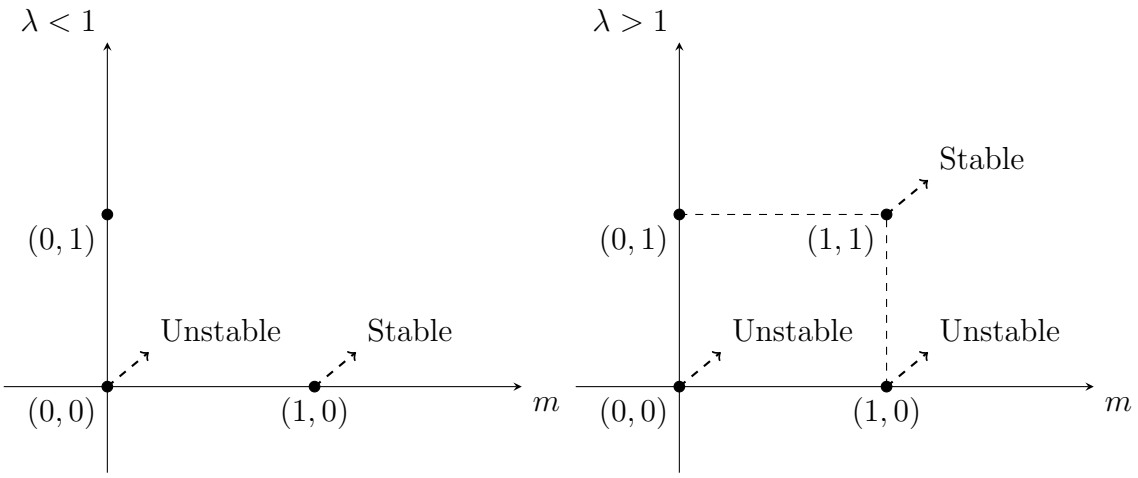


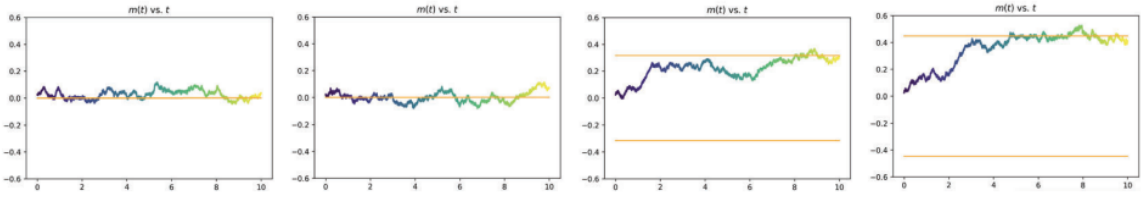
FIGURE 4.1. Phase Diagrams for BBP Transition

$$d\tilde{u}_2 = -8\tilde{u}_2(\tilde{u}_2^2 - 1)dt$$

We see that $r_\perp^2$ now solves an autonomous ODE which converges exponentially to 1 and the equation for $\tilde{u}_1$ behaves like

$$d\tilde{u}_1 = 4\tilde{u}_1(\lambda - 1)dt + 2\sqrt{2}dB_t$$

This is an OU process which is mean-reverting when $\lambda < 1$ and mean-repllent when $\lambda > 1$.



(a) $\lambda = 0.8$ (b) $\lambda = 0.9$ (c) $\lambda = 1.1$ (d) $\lambda = 1.2$

FIGURE 4.2. (Ben Arous, Gheissari, and Jagannath [2024]) Matrix PCA in dimension $N = 2000$

4.3. XOR Gaussian Mixture Model. Consider the problem of supervised learning for an XOR-type Gaussian mixture model in $\mathbb{R}^N$. Suppose that we are given i.i.d. samples of the form $Y = (\epsilon_l, Y_l)_l$, where y is Ber(1/2) and X has the following distribution,

$$Y_l = \begin{cases} \frac{1}{2}\left(\mathcal{N}(\mu, I) + \mathcal{N}(-\mu, I)\right), & \epsilon = 1 \\ \frac{1}{2}\left(\mathcal{N}(\nu, I) + \mathcal{N}(-\nu, I)\right), & \epsilon = 0 \end{cases}$$

where μ, ν are orthogonal unit vectors. Here, ϵ is the class label and Y is the data.

We focus on a two-layer neural network as a demonstration of our techniques for analyzing training dynamics on natural data models. This model, while simple, exhibits complex phenomenology.

Consider the classification problem using a two-layer neural network with our class label estimator defined as $\hat{\epsilon}(Y) = \sigma(v, g(WY))$, where σ is a sigmoid function and g is a ReLU function. Let W be a $K \times N$ matrix and v a K-vector. The loss function is given by:

$$L((v_i, W_i)_{i \le K}, (\epsilon, Y)) = -\epsilon \langle v, g(WY) \rangle + \log(1 + e^{\langle v, g(WY) \rangle}) + p(v, W)$$

where

$$p(v, W) := \frac{\alpha}{2}(\|v\|^2 + \|W\|^2).$$

The loss law at a point (v, W) depends only on the following variables for $1 \le i \le K$:

$$v_i, \quad m_i^\mu = \langle W_i, \mu \rangle, \quad m_i^\nu = \langle W_i, \nu \rangle.$$

If $\mathbf{u}_N$ is defined by these variables, the localizability criterion holds for fixed $\lambda > 0$ as long as $\delta_N = O(1/N)$. The main result can then be applied to obtain limits in both ballistic and diffusive phases.

In the large λ limit as $N \to \infty$, the effective dynamics in the ballistic phase (ODE system) become tractable. If $\alpha > 1/8$, the only fixed point is at $\mathbf{u}_N = \mathbf{0}$. However, if $0 < \alpha < 1/8$, the fixed points can be characterized by disjoint subsets $(I_0, I_\mu^+, I_\mu^-, I_\nu^+, I_\nu^-)$, with properties as follows (Ben Arous, Gheissari, and Jagannath [2024]):

(1) $m_i^\mu = m_i^\nu = v_i = 0$ for $i \in I_0$,
(2) $m_i^\mu = v_i > 0$ such that $\sum_{i \in I_\mu^-} v_i^2 = \text{logit}(-4\alpha)$ and $m_i^\nu = 0$ for all $i \in I_\mu^-$,
(3) $-m_i^\mu = v_i > 0$ such that $\sum_{i \in I_\mu^+} v_i^2 = \text{logit}(-4\alpha)$ and $m_i^\nu = 0$ for all $i \in I_\mu^+$,
(4) $m_i^\nu = v_i < 0$ such that $\sum_{i \in I_\nu^+} v_i^2 = \text{logit}(-4\alpha)$ and $m_i^\mu = 0$ for all $i \in I_\nu^+$,
(5) $-m_i^\nu = v_i < 0$ such that $\sum_{i \in I_\nu^-} v_i^2 = \text{logit}(-4\alpha)$ and $m_i^\mu = 0$ for all $i \in I_\nu^-$.

In the case of $K = 4$, these form 39 connected sets of fixed points, with $4! = 24$ being stable fixed points corresponding to possible permutations.

Since the derivations of the ballistic limiting equations apply for general K, we can study the probability of ballistic convergence to stable versus unstable fixed points as K varies. This is relevant for the XOR Gaussian Mixture Model (GMM), as $K = 4$ suffices for a Bayes-optimal classifier. The probability of being in the ballistic domain of attraction of stable fixed points is denoted by $p_{K,\lambda}$, with $p_{4,\lambda} \to \frac{3}{27}$ as $N \to \infty$ and $p_{K,\lambda} \to 1$ exponentially fast as $K \to \infty$, specifically:

$$p_k \asymp 1 - e^{-cK} \quad \text{for some } c > 0.$$

This indicates the advantages of overparameterization in a two-layer network. A random initialization is more likely to align with the "right" initial signature, ensuring that SGD will ballistically converge to a global minimizer of the population loss. This serves as a rigorous example of the well-known **lottery ticket hypothesis** (Frankle and Carbin [2018])

CHAPTER 13

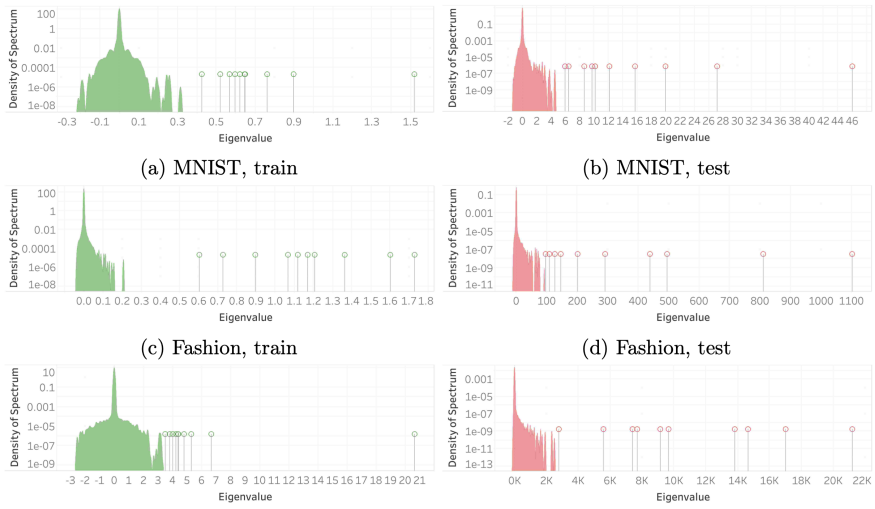
Spectral Alignment

1. Introduction

As happened in previous two lectures, we are interested in the interplay between training dynamics and the evolution of the spectra of natural geometric objects such as the Hessian and information matrices. In particular, our focus for today is going to be on how does the **local** geometry of the loss landscape is reflected along the training trajectory. (In physics terms, it is **the environment viewed from the particle**) About a decade ago, there is this conjecture/proposal by Yann Lecun LeCun et al. [2012] and Levent Sagun Sagun et al. [2017a,b] that: the landscape has many "flat directions", and SGD somehow "finds" the hidden low-dimensional structure.

More specifically, in terms of the quantitative measures on the spectra of the Hessian:

- There is a non-trivial bulk (of eigenvalues) centered at 0
- Other than the bulk, there is a few outlier that carries the structure of the training data (and underlying architecture)



(a) MNIST, train (b) MNIST, test

(c) Fashion, train (d) Fashion, test

(e) CIFAR10, train (f) CIFAR10, test

FIGURE 1.1. One of the numerical evidences in Papyan [2020], where the result is observed across a wide variety of architectures and learning tasks.

1.1. Set-up. As before, the general set-up is in supervised classification, where we are given data

$$\mathbf{Y} = (y, Y) \in \mathbb{R}^k \times \mathbb{R}^d$$

over distribution $\mathcal{P}_Y$, where $y \in \mathbb{R}^k$ is our label[1] that takes 1 on a class (from the index set $[k]$) and $Y \in \mathbb{R}^d$ is the corresponding feature vector. Our loss function is of the form

$$L(\mathbf{x}, \mathbf{Y}) : \mathbb{R}^p \times \mathbb{R}^{k \times d} \to \mathbb{R}_+$$

where $\mathbf{x} \in \mathbb{R}^p$ represents the network parameter (more specifically, it will be the usual cross-entropy loss corresponding to the architecture used). In addition, we have M training samples and $\widetilde{M}$ test samples[2], i.e., the training set $(\mathbf{Y}^l)_{l=1}^M$ and the test set $(\widetilde{\mathbf{Y}}^l)_{l=1}^{\widetilde{M}}$. To ensure SGD doesn't go off to infinity, we add an l^2 penalty term with Lagrange multiplier β.

Therefore, in summary, our online SGD with initialization $\mathbf{x}_0$ and learning rate δ, runs on the training set $(\mathbf{Y}^l)_{l=1}^M$ as follows:

$$\mathbf{x}_l = \mathbf{x}_{l-1} - \delta \nabla L(\mathbf{x}_{l-1}, \mathbf{Y}^l) - \beta \mathbf{x}_{l-1}$$

Our primary focus of concern is going to be understanding the spectra of the **empirical Hessian matrix** and **empirical second moment matrix** of the gradient, which are, respectively:

$$\nabla \hat{R}(\mathbf{x}) = \frac{1}{\widetilde{M}} \sum_{l=1}^{\widetilde{M}} \nabla^2 L(\mathbf{x}, \widetilde{Y}^l) \quad \text{and} \quad \hat{G}(\mathbf{x}) = \frac{1}{\widetilde{M}} \sum_{l=1}^{\widetilde{M}} \nabla L(\mathbf{x}, \widetilde{Y}^l)^{\otimes 2}$$

where the empirical risk is given by $\hat{R}(\mathbf{x}) = \frac{1}{\widetilde{M}} \sum_{l=1}^{\widetilde{M}} \nabla L(\mathbf{x}, \widetilde{Y}^l)$. Notice that $\hat{G}$ is exactly the (Fisher) information matrix when L is the log-likelihood. [3]

Specifically, the two cases we will be looking into is the Multi-class Logistic Regression and the XOR (Gaussian Mixture) problem. Informally, after a time roughly proportional to the dimensions:

- (These matrices has a block-diagonal structure, and) the eigenvalues of each block will largely concentrate near 0. However, there will be exactly C_k significant outliers, which doesn't depend on the number of communities in the tasks but instead the **hidden communities** within the problem.
- Online SGD aligns with its principal eigenspace
- In multilayer setting, alignment happens **within** each layer
- Even if the evolution starts to approach sub-optimal classifiers, the spectrum will also exhibit some degeneracy (i.e. alignment still occurs even near bad classifiers.)

We will now be diving into specific examples.

1.2. K-Gaussian Mixture Model (K-GMM). We have the following setup for the data model, let:

- $C \triangleq [k]$ be the collection of classes, where k is fixed.

[1] In class, our notation for the label is ϵ for the clarity of board writing, and here I am sticking with the notation in the paper.

[2] Just another discrepancy in the notation: in class, it is N training samples and M test samples.

[3] Again, just to recall, the reason we don't care about Hessian alone, is that 1. online SGD is a first-order method, and therefore Hessian may not contain enough information of the direction of descent (unless we are in a convex setting), and 2. calculating the Hessian might be computationally expensive. One benefit of the G-matrix is that it encompasses the average directions across possible moves, and its principal subspace tells you the main directions you'd want to take for a first-order method.

- $(\mu_a)_{a\in[k]} \in \mathbb{R}^d$ be the corresponding class means, with $\|u_a\| = 1$
- I_d/λ be the corresponding variances.
- $0 < (p_a)_{a\in[k]} < 1$ be the corresponding probabilities with $\sum_{a\in[k]} p_a = 1$

For the task to be solvable by **single-layer architecture**, we assume the means $(\mu_a)_{a\in[k]}$ are linearly independent, say with a fixed matrix of inner products $(\overline{m}_{ab})_{a,b} = (\mu_a \cdot \mu_b)_{a,b}$. Our data distribution $\mathcal{P}_Y$ is a mixture of the form $\sum_c p_c \mathcal{N}(\mu_c, I_d/\lambda)$, with the corresponding class label $y \in \mathbb{R}^k$. Namely, our data is given as $\mathbf{Y} = (y, Y)$ where

$$y \sim \sum_{a\in[k]} p_a \delta_{\mathbf{1}_a} \quad \text{and} \quad Y \sim \sum_{a\in[k]} y_a \mu_a + Z_\lambda$$

where $Z_\lambda \sim \mathcal{N}(0, I_d/\lambda)$.

$$L(\mathbf{x}, \mathbf{Y}) = -\sum_{c\in[k]} y_c x^c \cdot Y + \log \sum_{c\in[k]} \exp(x^c \cdot Y)$$

where $\mathbf{x} = (x^c)_{c\in C}$ are the parameters, each of which is a vector in $\mathbb{R}^d$, i.e. $\mathbf{x} \in \mathbb{R}^{dk}$ (We can alternatively view $\mathbf{x}$ as a $k \times d$ matrix)

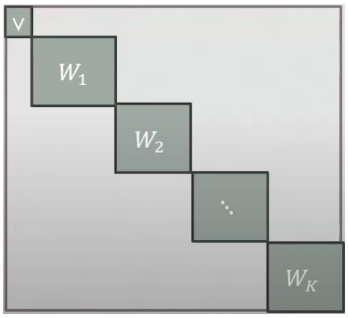


FIGURE 1.2. The principle block structure of Hessian and G-matrix we will be focusing on. Taken from Jagannath [2023]

There are two quantitative measures helping us the solidify the notion of 'alignment':

For a subspace B, we let P_B denote the orthogonal projection onto B; for a vector v, we let $\|v\|$ be its ℓ^2 norm; and for a matrix A, let $\|A\| = \|A\|_{op}$ be its $\ell^2 \to \ell^2$ operator norm.

DEFINITION 13.1. The *alignment* of a vector v with a subspace B is the ratio $\rho(v, B) = ||P_B v||/||v||$. We say a vector v *lives in a subspace* B up to error ε if $\rho(v, B) \geq 1 - \varepsilon$.

For a matrix A, we let $E_k(A)$ denote the span of the top k eigenvectors of A, i.e., the span of the k eigenvectors of A with the largest absolute values. We also use the following.

DEFINITION 13.2. We say a matrix A *lives in* a subspace B. up to error ε if there exists M such that $\mathrm{Im}(A - M) \subset B$ with $\|M\|_{op} \leq \varepsilon \|A\|_{op}$, where $\|A\|_{op}$ denotes the ℓ^2-to-ℓ^2 operator norm.

This definition will be made to quantitatively characterize the 'bulk' and 'outlier' mentioned before. Intuitively, it is saying that all eigenvectors with large eigenvalues live in the span of B, and all other eigenvalues are quantitatively small, up to ε-fraction in size.

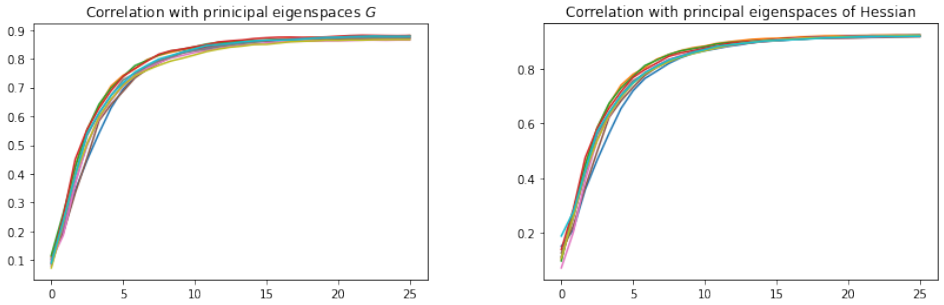
We are now ready to state the first result.

Informally: *after some linearly many steps, the SGD finds the subspace generated by the outlier eigenvalues of the Hessian and/or G-matrix of the test loss and lives there for future times.*

THEOREM 13.3. *Consider the mixture of k-Gaussians with cross-entropy loss function, and SGD with learning rate $\delta = O(1/d)$, regularizer $\beta > 0$, initialized from $\mathcal{N}(0, I_d/d)$. There exists α_0, λ_0 such that if $\lambda \geq \lambda_0$, and $\widetilde{M} \geq \alpha_0 d$, the following hold. For every $\varepsilon > 0$, there exists $T_0(\varepsilon)$ such that for any fixed time horizon $T_0 < T_f < M/d$, with probability $1 - o_d(1)$,*

$$\mathbf{x}_\ell^c \text{ lives in } E_k(\nabla^2_{cc}\widehat{R}(\mathbf{x}_\ell)) \text{ and in } E_k(\widehat{G}_{cc}(\mathbf{x}_\ell))\,,$$

for every $c \in [k]$, up to $O(\varepsilon + \lambda^{-1})$ error, for all $\ell \in [T_0\delta^{-1}, T_f\delta^{-1}]$.



(A) (B)

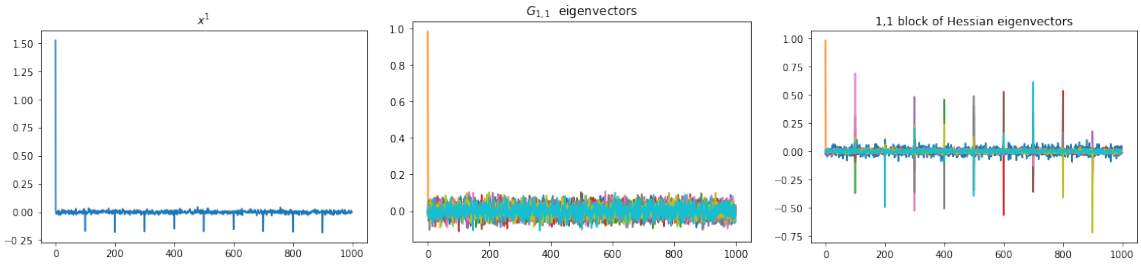
FIGURE 1.3. The alignment of the SGD trajectory $\mathbf{x}_\ell^c$ with $E_k(\nabla^2_{cc}\widehat{R}(\mathbf{x}_\ell))$ (left) and $E_k(\widehat{G}_{cc}(\mathbf{x}_\ell))$ (right), for $c \in [k]$. The x-axis is rescaled time, $\ell\delta$. The parameters are $k = 10$ classes in dimension $d = 1000$ with $\lambda = 10$, $\beta = 0.01$, and $\delta = 1/d$.

This result is demonstrated in Figure 1.3 which plots the alignment of the training dynamics $\mathbf{x}_\ell^c$ with the principal eigenspaces of the Hessian and G for each $c \in [k]$. As we see the alignment increases to near 1 rapidly for all blocks in both matrices. This theorem, and all our future results, are stated using a random Gaussian initialization, scaled such that the norm of the parameters is $O(1)$ in d. The fact that this is Gaussian is not relevant to the results, and similar results hold for other uninformative initializations with norm of $O(1)$.

Theorem 13.3 follows from the following theorem that describes the SGD trajectory, its Hessian (and its top k eigenspace), and its G-matrix (and its top k eigenspace), all live up to $O(1/\lambda)$ error in $\text{Span}(\mu_1, ..., \mu_k)$.

THEOREM 13.4. *In the setup of Theorem 13.3, the following live in $Span(\mu_1, ..., \mu_k)$ up to $O(\varepsilon + \lambda^{-1})$ error with probability $1 - o_d(1)$:*

1. *The state of the SGD along training, $\mathbf{x}_\ell^c$ for every c;*
2. *The b, c blocks of the empirical test Hessian, $\nabla^2_{bc}\widehat{R}(\mathbf{x}_\ell)$ for all $b, c \in [k]$;*
3. *The b, c blocks of the empirical test G-matrix $\widehat{G}_{bc}(\mathbf{x}_\ell)$ for all $b, c \in [k]$.*



(A) (B) (C)

FIGURE 1.4. From left to right: Plot of entries of $\mathbf{x}_\ell^1$ and the k leading eigenvectors of $\nabla_{11}^2 \widehat{R}(\mathbf{x}_\ell)$ and $\widehat{G}_{11}(\mathbf{x}_\ell)$ respectively at the end of training, namely $\ell = 50 \cdot d = 25,000$ steps. Here the x-axis represents the coordinate. The parameters are the same as in Fig. 1.3 and the means are $\mu_i = e_{i*50}$.

We demonstrate this result in Figure 1.4, which shows the coordinate-wise values of a fixed block of the SGD, the Hessian, and the G-matrix.

Inside the low-rank space spanned by $\mu_1, ..., \mu_k$, the training dynamics, Hessian and G-matrix spectra can display different phenomena depending on the relative locations of $\mu_1, ..., \mu_k$ and weights $p_1, ..., p_k$. To illustrate the more refined alignment phenomena, let us take as a concrete example $p_c = \frac{1}{k}$ for all c, and $\mu_1, ..., \mu_k$ orthonormal.

With this concrete choice, we can analyze the limiting dynamical system of the SGD without much difficulty and its relevant dynamical observables have a single stable fixed point, to which the SGD converges in linearly many, i.e., $O(\delta^{-1})$, steps. This allows us to show more precise alignment that occurs within $\text{Span}(\mu_1, ..., \mu_k)$ over the course of training.

THEOREM 13.5. *In the setting of Theorem 13.3, with the means $(\mu_1, ..., \mu_k)$ being orthonormal, the estimate $\mathbf{x}_\ell^c$ has $\Omega(1)$, positive, inner product with the top eigenvector of both $\nabla_{cc}^2 \widehat{R}(\mathbf{x}_\ell)$ and $\widehat{G}_{cc}(\mathbf{x}_\ell)$ (and negative, $\Omega(1)$ inner product with the $k-1$ next largest eigenvectors). Also, the top eigenvector of $\nabla_{cc}^2 \widehat{R}(\mathbf{x}_\ell)$, as well as that of $\widehat{G}_{cc}(\mathbf{x}_\ell)$, live in Span$(\mu_c)$ up to $O(\varepsilon + \lambda^{-1})$ error.*

Just to summarize: the above three results tell us the 3 aspects of the training dynamics (for the classification of k-GMM):

- At initialization, and throughout the parameter space in each class block, the Hessian and G-matrices decompose into a rank-k outlier part spanned by $\mu_1, \cdots, \mu_k$, and a correction term of size $O(1/\lambda)$ in operator norm.
- Furthermore, when initialized randomly, the SGD is not aligned with the outlier eigenspace, but does align with them in a short $O(\delta^{-1})$ number of steps.
- Moreover, when the means are orthogonal, each class block of the SGD $\mathbf{x}_l^c$ in fact correlates strongly with the specific mean for its classes μ_c, and simultaneously, in the Hessian and G-matrices along training, the eigenvalue corresponding to μ_c becomes distinguished from the other $k-1$ outliers.

1.3. XOR-Gaussian Mixture Model. We consider this more complex XOR model which requires a multi-layer architecture to solve. Specifically, we consider the problem of classifying a 4-component Gaussian mixture with means $\pm\mu$ and $\pm\nu$, where $\|\mu\| = \|\nu\| = 1$.

For the sake of simplicity, we assume the means are orthogonal with variance I_d/λ. There are two classes: class label 1 for Gaussian with mean $\pm\mu$ and 0 for Gaussian with mean $\pm\nu$. More precisely:

$$y \sim \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1 \quad \text{and} \quad Y \sim \begin{cases} \frac{1}{2}\mathcal{N}(\mu, I_d/\lambda) + \frac{1}{2}\mathcal{N}(-\mu, I_d/\lambda) & y = 1 \\ \frac{1}{2}\mathcal{N}(\nu, I_d/\lambda) + \frac{1}{2}\mathcal{N}(-\nu, I_d/\lambda) & y = 0 \end{cases}$$

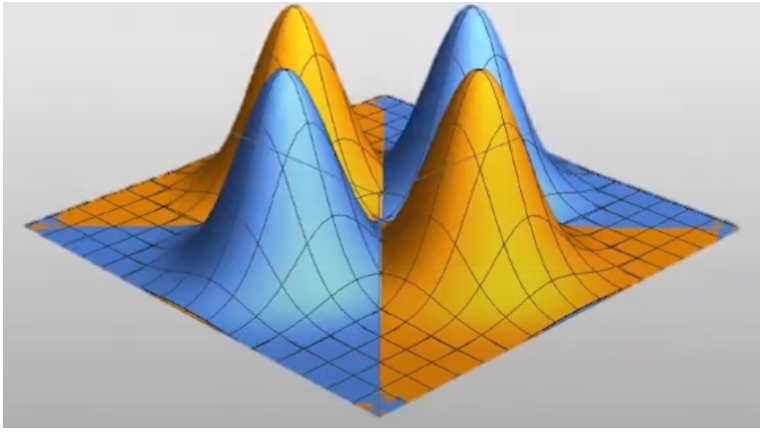


FIGURE 1.5. Visualization of a mixture of 4 Gaussian, taken from Jagannath [2023]

To solve this, we need:

- A two-layer NN with intermediate layer having at least width $K \geq 4$
- ReLU activation function $g(x) = x \vee 0$
- The sigmoid activation function $\sigma(x) = \frac{1}{1+e^{-x}}$
- The parameter space $\mathbf{x} = (W, \nu) \in \mathbb{R}^{Kd+K}$, where $W(\mathbf{x}) \in \mathbb{R}^{K \times d}$ is the first-layer weights and $\nu(bx) \in \mathbb{R}^K$ is the second layer weight.
- A binary cross-entropy loss of the form:

$$L(\mathbf{x}, \mathbf{Y}) = -yv \cdot g(WY) + \log(1 + e^{v \cdot g(WY)})$$

with g applied entry-wise.

It turns out a similar story (as in the case of 1-layer NN on k-GMM) can happen for this model problem.

THEOREM 13.6. *Consider the XOR GMM mixture with loss function* (**??**) *and the corresponding SGD* (**??**) *with* $\beta \in (0, 1/8)$*, learning rate* $\delta = O(1/d)$*, initialized from* $\mathcal{N}(0, I_d/d)$*. There exist* α_0, λ_0 *such that if* $\lambda \geq \lambda_0$*, and* $\widetilde{M} \geq \alpha_0 d$*, the following hold. For every* $\varepsilon > 0$*, there exists* $T_0(\varepsilon)$ *such that for any fixed time horizon* $T_0 < T_f < M/d$*, with probability* $1 - o_d(1)$*, for all* $i \in \{1, ..., K\}$*,*

(1) $W_i(\mathbf{x}_\ell)$ *lives in* $E_2(\nabla^2_{W_i W_i} \widehat{R}(\mathbf{x}_\ell))$ *and in* $E_2(\widehat{G}_{W_i W_i}(\mathbf{x}_\ell))$*, and*

(2) $v(\mathbf{x}_\ell)$ *lives in* $E_4(\nabla^2_{vv} \widehat{R}(\mathbf{x}_\ell))$ *and* $E_4(\widehat{G}_{vv}(\mathbf{x}_\ell))$*,*

up to $O(\varepsilon + \lambda^{-1/2})^4$ *error, for all* $\ell \in [T_0 \delta^{-1}, T_f \delta^{-1}]$.

Note the restriction to $\beta < 1/8$ in Theorems 13.6–13.7 is because when $\beta > 1/8$ the regularization is too strong for the SGD to be meaningful;

This theorem is demonstrated in Figure 1.6. There we have plotted the alignment of the rows in the intermediate layer with the space spanned by the top two eigenvectors of the corresponding first-layer blocks of the Hessian and G-matrices, and similarly for the final layer.

As before, the above theorem follows from the following theorem that describes both the SGD trajectory, its Hessian, and its G-matrix, living up to $O(\varepsilon + \lambda^{-1/2})$ error in their first-layer blocks in $\mathrm{Span}(\mu, \nu)$ and in their second layer blocks in

$$\mathrm{Span}\{\ g(W(\mathbf{x}_\ell)\mu),\ g(-W(\mathbf{x}_\ell)\mu),\ g(W(\mathbf{x}_\ell)\nu),\ g(-W(\mathbf{x}_\ell)\nu))\}$$

where g is applied entry-wise.

THEOREM 13.7. *In the setting of Theorem 13.6, up to* $O(\varepsilon + \lambda^{-1/2})$ *error with probability* $1 - o_d(1)$*, the following live in* $\mathrm{Span}(\mu, \nu)$*,*

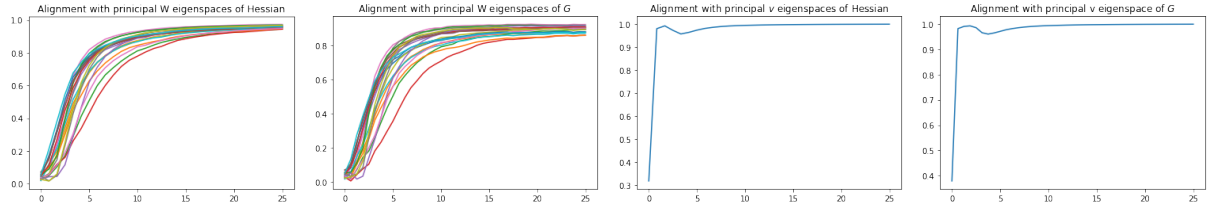
- *The first layer weights,* $W_i(\mathbf{x}_\ell)$ *for each* $i \in \{1, ..., K\}$*,*
- *The first-layer empirical test Hessian* $\nabla^2_{W_i W_i} \widehat{R}(\mathbf{x}_\ell)$ *for each* $i \in \{1, ..., K\}$*,*
- *The first-layer empirical test G-matrix* $\widehat{G}_{W_i W_i}(\mathbf{x}_\ell)$ *for each* $i \in \{1, ..., K\}$*,*

and the following live in $\mathrm{Span}(g(W(\mathbf{x}_\ell)\mu), g(-W(\mathbf{x}_\ell)\mu), g(W(\mathbf{x}_\ell)\nu), g(-W(\mathbf{x}_\ell)\nu))$

- *The second layer weights* $v(\mathbf{x}_\ell)$*,*
- *The second-layer empirical test Hessian* $\nabla^2_{vv} \widehat{R}(\mathbf{x}_\ell)$*,*

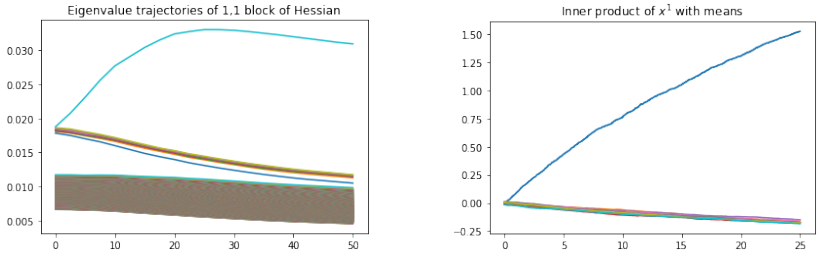
[4] In this theorem and in Theorem 13.7, the big-O notation also hides constant dependencies on the initial magnitudes of $v(\mathbf{x}_0)$, and on T_f.

- *The second-layer empirical test G-matrix* $\widehat{G}_{vv}(\mathbf{x}_\ell)$.



(A) (B) (C) (D)

FIGURE 1.6. (a) and (b) depict the alignment of the first layer weights $W_i(\mathbf{x}_\ell)$ with the principal subspaces of the corresponding blocks of the Hessian and G-matrices, i.e., with $E_2(\nabla^2_{W_iW_i}\widehat{R}(\mathbf{x}_\ell))$ and $E_2(\widehat{G}_{W_iW_i}(\mathbf{x}_\ell))$. (c) and (d) plot the second-layer alignment, namely of $v(\mathbf{x}_\ell)$ with $E_4(\nabla^2_{vv}\widehat{R}(\mathbf{x}_\ell))$ and $E_4(\widehat{G}_{vv}(\mathbf{x}_\ell))$. Parameters are $d = 1000$, $\lambda = 10$, and $K = 20$.

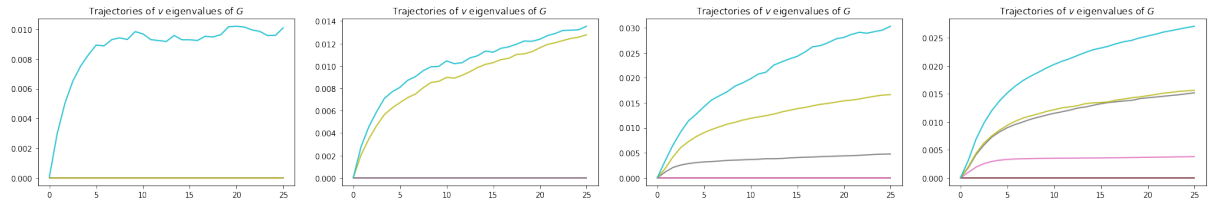


(A) (B)

FIGURE 1.7. Left: the eigenvalues of $\nabla^2\widehat{R}_{11}(\mathbf{x}_\ell)$ over the course of training. The leading k eigenvalues are separated from the bulk at all times, and the top eigenvalue, corresponding to μ_1, separates from the remaining eigenvalues soon after initialization. Right: the inner product of $\mathbf{x}^1_\ell$ with the means $\mu_1, ..., \mu_k$ undergoes a similar separation over the course of training. Parameters are the same as in preceding figures.

1.4. Further Directions and Open Questions.

- A more precise understanding of **Effective Spectral Theory**: extend the analysis of training dynamics of NN on other problems (using similar analysis on the Hessian and G-matrices); Moreover, while the effective spectrum has primarily been studied in the context of **online learning**, exploring its role in **batch learning** could uncover new optimization strategies. In particular, batch learning introduces dependencies across samples, which may influence spectral properties differently from online settings. Analyzing these dependencies could lead to refined sampling strategies for batch training that emphasize high-variance directions, enhancing convergence rates. Another direction is to establish a deeper connection with **BBP Phase Transition** in Neural Networks. Further, we can also investigate Effective Spectral



(A) (B) (C) (D)

FIGURE 1.8. Evolution of eigenvalues in the v component of G over time in rank-deficient cases. Here, SGD is started from initializations that converge to suboptimal classifiers (this has uniformly positive, K-dependent, probability under a random initialization). From left to right, the SGD's classifier varies in the number of hidden classes it discerns, from 1 to 4. There is still a dynamical spectral transition, now with only a corresponding number of emerging outlier eigenvalues.

Theory Beyond Gaussian Regime (e.g. Heavy tail distributions, which also occurs quite frequently in practice)

- A spectrally adapted version of **AdaGrad** and **RMSprop**: Such an algorithm would modify learning rates based on the spectral characteristics of each layer, dynamically adjusting to ensure that the most critical directions are optimized first. This kind of spectral adaptation could address the common problem of "plateauing" during training, where progress stalls due to poor alignment between the learning rate and the underlying landscape geometry.
- Offer some explanations on the **knowledge transfer** across layers. (In a similar manner, we may want to identify the principal subspaces that each layer aligns with, and doing so might be able to share information between layers in a way that preserves the structural coherence of the model. For instance, rather than training each layer independently, one could guide each layer's learning process by aligning it with the subspaces learned by the preceding layer. This approach could allow for faster training and greater parameter efficiency, especially in over-parameterized models.)
- Implications for **Generalization and Robustness**

2. Summary of technicalities and proofs

Here will informally summarize further technicality and machineries used in the proofs. The organization Ben Arous et al. [2023a] is as follows:

- Section 3-4 quantitatively characterizes the 'bulks' of the population Hessian and G-matrices. Section 3 deals with the k-GMM problem (1-layer NN) and Section 4 deals with the XOR problem (2-layer NN), specifically:
  - **Lemma 3.1.** calculates the off-diagonal blocks of the Hessian (in expectation)
  - **Lemma 3.2.** calculates the on-diagonal blocks of the Hessian (in expectation)
  - **Lemma 3.3.** calculates the sub-blocks of the population G-matrix.

- **Lemma 4.1.** concentration of the ReLU function g, in the form:
$$\|\mathbb{E}[g(WX)^{\otimes 2} - g(W\vartheta)] \lesssim_K \psi_2(\vartheta, W, \lambda)$$
for $X \sim \mathcal{N}(\vartheta, I_d/\lambda)$ and W being $K \times d$
- **Lemma 4.2.** bound on $\mathbb{E}[\|g(WX)^{\otimes 2}\|_{\text{op}}^2]^{1/2}$
- **Lemma 4.3.** bound on the expectations on the composition of σ, σ' with $\vartheta \cdot g(WX)$
- **Lemma 4.4.** concentration/bounds on the Hessian (on the first layer block $(v, W) \in B$), of the form
$$\left\| \mathbb{E}[\nabla_{vv}^2 L] - 1/4 \sum_{\vartheta \in \{\pm\mu, \pm\nu\}} \sigma'(v \cdot g(W\nu)) \cdot g(W\vartheta)^{\otimes 2} \right\|_{op} \lesssim \varphi_2(\nu, W, \lambda) + \frac{1}{\lambda}$$
- **Lemma 4.5.** concentration/bound on the G-matrix (of the first layer block), of the form:

$$||\mathbb{E}[\nabla_v L^{\otimes 2}] - \frac{1}{4} \sum_{\vartheta \in \{\pm\mu, \pm\nu\}} (y_\vartheta - \sigma(v \cdot g(W\vartheta)))^2 g(W\vartheta)^{\otimes 2}|| \lesssim_{K,B} \psi_2(\vartheta, W, \lambda) + \frac{1}{\lambda},$$

and a similar bound on second layer blocks.

- Section 5 analyzes the SGD-trajectory for k-GMM and XOR problems, using the effective dynamics framework in Ben Arous et al. [2023b] (of course with some twisting of the original statements and arguments)
  - **Prop 5.1.** (Informal: For k-GMM) The weights $\|\mathbf{x}_l^c\|$ are bounded after enough number of SGD iterations, and $\mathbf{x}_l^c$ is within $O(\varepsilon + \lambda^{-1})$ distance of a point in $\text{Span}\{\mu_1, \cdots, \mu_k\}$ having $\|\mathbf{x}^c\| > \eta$
  - **Prop 5.2.** An analogue of Prop 5.1. for XOR-GMM
  - **Def 5.3 - Lem 5.6.**: Recalling effective dynamics/summary statistics.
  - **Thm 5.7 - Cor 5.12**: Summary statistics adaptation for our case problems.
  - **Lem 5.13 - Lem 5.22**: Analysis of the subspace generated by the means.
- Section 6 proves various concentration results for empirical Hessian and G-matrices
- Section 7 proves all of major theorems using building blocks from Section 3 - 6

CHAPTER 14

Transformers

1. Introduction

The goal is to develop a mean-field/interacting-particle system perspective on the mechanism of **Transformers**. From a high-level point of view: transformers are trained to predict the next token. The ordering of the input token, due to positional encoding [Vaswani et al. [2017]], does not matter. Therefore we can have an empirical measure on the set of tokens, and hence the input can be seen as the probability measures on tokens. The similar view point can be carried to the output space (which is just the likelihood of token being predicted next). Therefore transformers can be viewed as some parametrized **flow-maps** between the set of probability measures on Euclidean spaces; and since they operate on the input token 'collectively', we can then have some governing PDE that describes the collective dynamics.

1.1. A warm-up: ResNet. Again, in supervised learning, we are interested in approximating a model $f : \mathbb{R}^d \to \mathbb{R}^m$ from a family of empirical data $\mathcal{D} = \{x^{(i)}, f(x^{(i)}\}_{i\in[N]}$, and ResNet is one of the models for learning f. In its basic form:

$$\begin{cases} x(0) = x \\ x(k+1) = x(k) + w(k)\sigma(a(k)x(k) + b(k)) \quad \text{for } k \in \{0, \cdots, L-1\} \end{cases} \tag{1.1}$$

which has a total of $L+1$-layers (with L-hidden layers). Here σ is a Lipschitz function applied component-wise to the input vector, while $\theta(\cdot) = (w(\cdot), a(\cdot), b(\cdot)) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d} \times \mathbb{R}^d$ are the parameters we wanted to train. The output $x_i(L)$ serving as a representation of $x^{(i)}$ is then fed into the last layer (which just corresponds to a classical learning task such as linear/logistic regression for predicting the label $f(x^{(i)})$. The key element here is the **skip connections**, meaning the previous step $x_i(k)$ appear explicitly in the iteration for the next one.

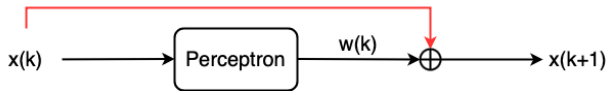


FIGURE 1.1. A simplified scheme of ResNet, the red arrow denots the 'skip connection'

1.2. Neural ODE. Note that the index k can be viewed as a time variable, and therefore we have a continuous-time analogue, **Neural Ordinary Differential Equation**, introduced as studied in Chen et al. [2018], E [2017], Haber and Ruthotto [2017]

$$\begin{cases} \dot{x}(t) = w(t)\sigma(a(t)x(t) + b(t)) \quad \text{for } t \in (0, T) \\ x(0) = x \end{cases} \tag{1.2}$$

1.3. Transformer as interacting particle system. In ResNet, the input is a single vector $x(0) \in \mathbb{R}^d$. In language tasks, Transformers operate on a sequence of vectors of length N, namely, $\{x_i(0)\}_{i\in[N]} \in (\mathbb{R}^d)^N$, where each vector $x_i(0) \in \mathbb{R}^d$ is a word, or a **token** formally, and the entire sequence $\{x_i(0)\}_{i\in[n]}$ is called a **prompt**. Most of practical implementation of transformers utilizes **layer normalization**, and so the input therefore has to be normalized, and hence $\{x_i(0)\}_{i\in[N]}$ is processed so that $x_i(0) \in \mathbb{S}^{d-1}$ for all $i \in [N]$. Therefore, to pursue a continuous viewpoint, we view a transformer as a flow-map on $(\mathbb{S}^{d-1})^N$ with $\{x_i(0)\}_{i\in[N]}$ as initial conditions evolving according to the dynamics:

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\frac{1}{Z_{\beta,i}(t)} \sum_{j=1}^{n} e^{\beta\langle Q(t)x_i(t), K(t)x_j(t)\rangle} V(t)x_j(t) \right) \tag{1.3}$$

for all $i \in [N]$ and $t \geq 0$, where $\mathbf{P}^{\perp}_x y = y - \langle x, y\rangle x$ denotes the projection of $y \in \mathbb{R}^d$ onto $T_x\mathbb{S}^{d-1}$, and the partition function $Z_{\beta,i}(t) > 0$ is

$$Z_{\beta,i}(t) = \sum_{k=1}^{n} e^{\beta\langle Q(t)x_i(t), K(t)x_k(t)\rangle} \tag{1.4}$$

where $(Q(\cdot), K(\cdot), V(\cdot))$ are parameters matrices learned from matrices [1], and β can be seen as "inverse temperature" from statistical physics. In layman's term: this is a parameter that controls the "sharpness" or "flatness" of a probability distribution. [2]

Statistical Physics	**Quasi-Bayesian Inference**
Energy Function (Hamiltonian/Lagrangian)	Log-Likelihood Function
State Space	Parameter Space
Inverse Temperature	Inverse Temperature [3]
Micro-canonical Ensemble	Prior

TABLE 1. Some terminology correspondences

A simplified version of (1.3)–(1.4) is first derived in Lu et al. [2020], Dutta et al. [2021], Sander et al. [2022]. In which the **self-attention** mechanism:

$$A_{ij}(t) = \frac{e^{\beta\langle Q(t)x_i(t), K(t)x_j(t)\rangle}}{Z_{\beta,t}} \quad (i,j) \in [n]^2$$

[1] Just an extremely concise explanation of what the Q, K, V matrices mean: essentially they are different projections of the same input. Q represents what the current token is "looking for" or asking about; K represents what information each token "contains"; V contains the actual information to be aggregated

[2] In class, we've also mentioned the the asymptotic behavior of Gibbs measure $\exp(-\beta E(x))/Z(\beta)$ when β is in high-temperature and low-temperature regime. Notice that 1.3 is just the analogue of the Gibbs measure

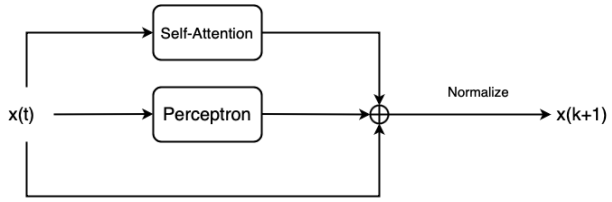


FIGURE 1.2. An extremely simplified scheme of transformer

is viewed a the **nonlinear coupling system** in the interacting particle system. The matrix A [4] is a $n \times n$ stochastic matrix (with rows being probability vectors). The reason behind the name **attention** is due to the fact that $A_{ij}(t)$ captures the attention given by particle i to particle j relatively to all particles $l \in [n]$. The neighbors for which a single particle pays attention to is determined by the $Q(t)$ and $K(t)$ matrices.

In addition, due to the observation of **clustering** of $\{\langle x_i(t), x_j(t)\rangle\}_{(i,j)\in[n]^2, i\neq j}$, and for the simplicity of our analysis, we will work with the case when $Q = K = V = \lambda I_d$. In the case when $lambda = 1$, the dynamics is:

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\frac{1}{Z_{\beta,i}(t)} \sum_{j=1}^{n} e^{\beta\langle x_i(t), x_j(t)\rangle} x_j(t) \right) \tag{SA}$$

for $i \in [n]$ and $t \geq 0$ and, as before

$$Z_{\beta,i}(t) = \sum_{k=1}^{n} e^{\beta\langle x_i(t), x_k(t)\rangle}. \tag{1.5}$$

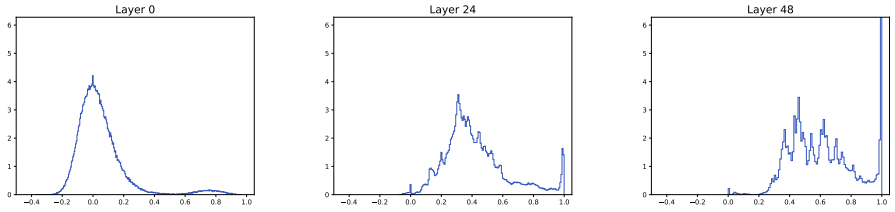


FIGURE 1.3. A subset of figures adapted from the original paper Geshkovski et al. [2024], demonstrating the **clustering** phenomenon: Histogram of $\{\langle x_i(t), x_j(t)\rangle\}_{(i,j)\in[n]^2, i\neq j}$ at different layers t in the context of the trained ALBERT XLarge v2 model (Lan et al. [2020] and https://huggingface.co/albert-xlarge-v2)[5], which has constant parameter matrices.

[4] In class, we used the notation SA for the self-attention matrix

1.4. A complete model with multi-headed attention. For practical purposes, the implementation of self-attention usually spreads out the computation of multiple-heads at every time, which is called **multi-headed attention**, and it leads to the following modification of the

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\sum_{h=1}^{H} \sum_{j=1}^{n} \frac{e^{\beta \langle Q_h(t)x_i(t), K_h(t)x_j(t) \rangle}}{Z_{\beta,i,h}(t)} V_h(t)x_j(t) \right) \tag{1.6}$$

where $H \geq 1$ here is the number of heads, and $Z_{\beta,i,h}(t)$ is defined as in (1.4) for the matrices $Q_h(t)$ and $K_h(t)$.

Lastly, we wanted to combine the the feed-forward layer, which leads to

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\sum_{h=1}^{H} \sum_{j=1}^{n} \frac{e^{\beta \langle Q_h(t)x_i(t), K_h(t)x_j(t) \rangle}}{Z_{\beta,i,h}(t)} V_h(t)x_j(t) + w(t)\sigma(a(t)x_i(t) + b(t)) \right), \tag{1.7}$$

where $w(t), a(t), b(t)$ and σ are all as in (1.2). The interested reader is referred to Lu et al. [2020], Phuong and Hutter [2022] for all the details[6].

2. The master equation and continuity-based analysis

As we've said in the introductory remarks, the input token sequence is encoded as a **set** of sequence $\{x_1(0), \cdots, x_n(0)\}$, and this can be seen as the empirical measure of its constituent tokens $1/n \sum_{i=1}^{n} \delta_{x_i(0)}$, and the output can hence also be seen as a probability measure in the form $1/n \sum_{i=1}^{n} \delta_{x_i(t)}$. Therefore, the transformer can be viewed as a flow map between probability measures on $\mathbb{S}^{d-1}$, and to describe this flow map, some governing equations has to be laid out.

The vector field driving the evolution of a single particle will depend on all n particles, and therefore, we can write the dynamics (SA) as

$$\dot{x}_i(t) = \mathcal{X}[\mu(t)](x_i(t)) \tag{2.1}$$

for all $i \in [n]$ and $t \geq 0$, where

$$\mu(t, \cdot) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i(t)}(\cdot)$$

is the empirical measure, while the vector field $\mathcal{X}[\mu] : \mathbb{S}^d \to T\mathbb{S}^{d-1}$ reads

$$\mathcal{X}[\mu](x) = \mathbf{P}^{\perp}_{x} \left(\frac{1}{Z_{\beta,\mu}(x)} \int e^{\beta \langle x,y \rangle} y d\mu(y) \right) \tag{2.2}$$

with

$$Z_{\beta,\mu}(x) = \int e^{\beta \langle x,y \rangle} d\mu(y). \tag{2.3}$$

[6] and lines 123–130 in OpenAI [2024] for some relevant source code.

In other words, (SA) is a *mean-field interacting particle system*. The evolution of $\mu(t)$ is governed by the continuity equation[7]

$$\begin{cases} \partial_t \mu + \mathtt{div}(\mathcal{X}[\mu]\mu) = 0 & \text{on } \mathbb{R}_{\geq 0} \times \mathbb{S}^{d-1} \\ \mu_{|t=0} = \mu(0) & \text{on } \mathbb{S}^{d-1} \end{cases} \tag{2.4}$$

satisfied in the sense of distributions.

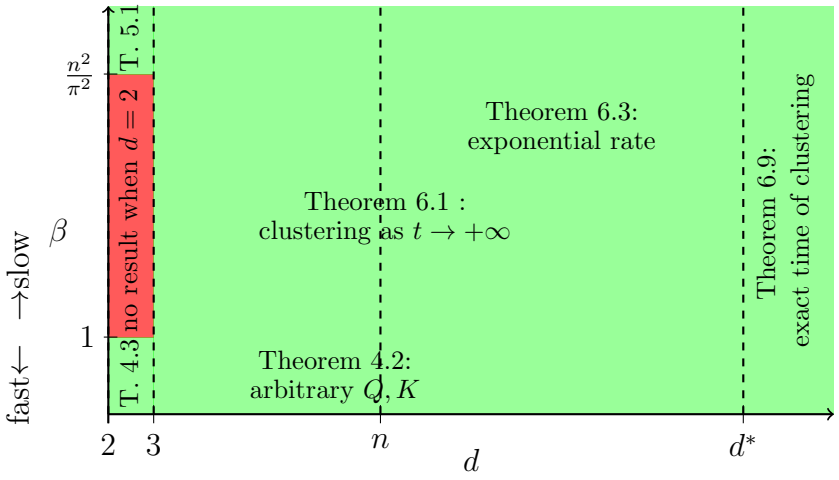


FIGURE 2.1. Green zones indicate regimes where convergence to a single cluster as $t \to +\infty$ can be proven. Here $n \geq 2$ is fixed. When d is larger than specific thresholds, the long-time asymptotics can be chiseled out in finer detail. Convergence is slow when $\beta \gg 1$ (relative to the size of d, n), as even the exponential decay constant when $d \geq n$ is of the form $\lambda = O(e^{-\beta})$ and thus degenerates.

2.1. The role of clusters and its implications. As demonstrated in the figure, the dimension d and β play a key role in the clustering results. Hereby we investigate the boundary cases of β while keeping d and n fixed.

In the case when $\beta = 0$, the continuity equation (SA) reads

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\frac{1}{n} \sum_{j=1}^{n} x_j(t) \right) \quad t \geqslant 0. \tag{2.5}$$

The first main result roughly says that over the initial points, a single cluster emerges.

THEOREM 14.1. *Let $d, n \geq 2$. For Lebesgue almost any initial sequence $(x_i(0))_{i\in[n]} \in (\S)^n$, there exists some point $x^* \in \mathbb{S}^{d-1}$ such that the unique solution $(x_i(\cdot))_{i\in[n]} \in C^0(\mathbb{R}_{\geq 0}; (\mathbb{S}^{d-1})^n)$ to the corresponding Cauchy problem for (2.5) satisfies*

$$\lim_{t\to+\infty} x_i(t) = x^*$$

for any $i \in [n]$.

[7]Unless stated otherwise, ∇ and $\mathtt{div}$ henceforth stand for the spherical gradient and divergence respectively, and all integrals are taken over $\mathbb{S}^{d-1}$.

The proof of this relies on standard arguments from dynamical system. The evolution given by 2.5 is gradient ascent for the energy function $\mathsf{E}_0 : (\mathbb{S}^{d-1})^n \to \mathbb{R}$ given by

$$\mathsf{E}_0(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle x_i, x_j \rangle.$$

Since the dynamics are the gradient ascent of a real-analytic functional on the compact real-analytic manifold $(\mathbb{S}^{d-1})^n$, the celebrated Łojasiewicz theorem Lojasiewicz [1963], in the form given by [Ha et al., 2018, Corollary 5.1]—which is valid in the context of general compact Riemannian manifolds—, implies that for any initial condition $X \in (\mathbb{S}^{d-1})^n$, the solution $\Phi^t(X) \in (\mathbb{S}^{d-1})^n$ converges to some critical point $X^* \in (\mathbb{S}^{d-1})^n$ of E_0 as $t \to +\infty$.

Formally, a **strict saddle point** of E_0 is a critical point of E_0 at which the Hessian of E_0 has at least one strictly positive eigenvalue. Then 14.1 follows from Łojasiewicz theorem with the following lemmas.

LEMMA 14.2. *Let $\mathcal{M}$ be a compact Riemannian manifold and let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function. The set of initial conditions $X_0 \in \mathcal{M}$ for which the gradient ascent*

$$\begin{cases} \dot{X}(t) = \nabla f(X(t)) \\ X(0) = X_0 \end{cases} \tag{2.6}$$

converges to a strict saddle point of f is of volume zero.

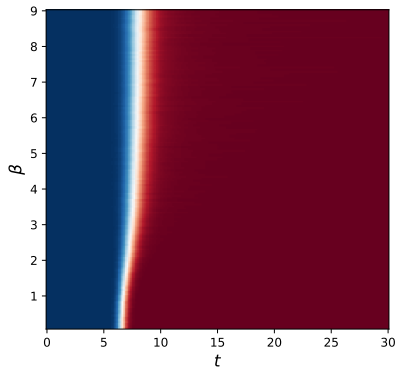
LEMMA 14.3. *Any critical point $(x_1, \dots, x_n) \in (\mathbb{S}^{d-1})^n$ of E_0 which is not a global maximum, namely such that $x_1 = \dots = x_n$, is a strict saddle point. In particular, all local maxima are global.*

The other major results deals with β in other regimes, and the similar results with 14.1 holds with mild conditions.

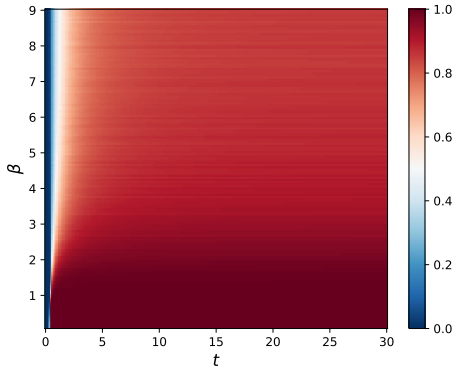
2.2. Some open questions and numerics. Below is some numerical experiments demonstrating the phase transition of the dynamics, relating to some open questions in their analysis.

Open question 1: Do the dynamics enter a transient metastable state, in the sense that for $\beta \gg 1$, all particles stay in the vicinity of $m < n$ clusters for long periods of time, before they all collapse to the final cluster $\{x^*\}$?

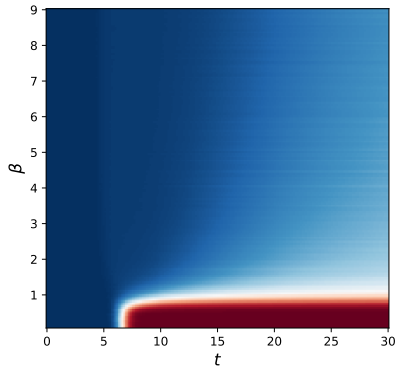
Open question 2 (informal): Can our major results hold with (Q, K, V) being random matrices?



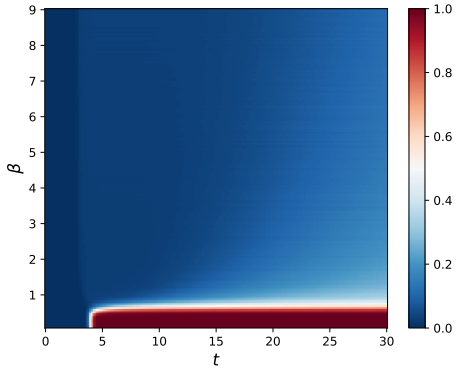
(A) Q, K, V in real Ginibre ensemble



(B) $Q^\top K$ Wigner, $V \succeq 0$ GOE



(C) Q, K in real Ginibre ensemble, $V = Q^\top K$



(D) $Q^\top K$ Wigner, $V = Q^\top K$

FIGURE 2.2. Phase diagrams for some choices of random matrices (Q, K, V); here $d = 128$, $n = 32$. Sharp phase transitions as well as metastable regions appear in all cases. In the gradient flow case $Q^\top K = V$, there is a resemblance to the well-understood case $Q^\top K = V = I_d$, but with an appearance of a metastable zone even in high dimension, which we are not able to explain for the moment.

Bibliography

Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294, 1993. ISSN 00911798, 2168894X. URL http://www.jstor.org/stable/2244575.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1611–1619. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/belkin19a.html.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.

Gerard Ben Arous, Reza Gheissari, Jiaoyang Huang, and Aukosh Jagannath. High-dimensional sgd aligns with emerging outlier eigenspaces, 2023a. URL https://arxiv.org/abs/2310.03010.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling, 2023b. URL https://arxiv.org/abs/2206.04030.

Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. *Communications on Pure and Applied Mathematics*, 77(3):2030–2080, 2024. ISSN 1097-0312. doi: 10.1002/cpa.22169. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.22169.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Subhabrata Dutta, Tanya Gautam, Soumen Chakrabarti, and Tanmoy Chakraborty. Redesigning the transformer architecture with insights from multi-particle dynamical systems. *Advances in Neural Information Processing Systems*, 34:5531–5544, 2021.

Weinan E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 1(5):1–11, 2017.

Jonathan Frankle and Michael Carbin. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. September 2018.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers, 2024. URL https://arxiv.org/abs/2312.10794.

Seung-Yeal Ha, Dongnam Ko, and Sang Woo Ryoo. On the relaxation dynamics of Lohe oscillators on some Riemannian manifolds. *Journal of Statistical Physics*, 172:1427–1478, 2018.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34(1), 2017.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Aukosh Jagannath. Spectral alignment for high dimensional sgd - vector institute machine learning theory workshop, 2023. URL `https://www.youtube.com/watch?v=8E0c7KX6zEU&t=403s`.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*, 2020.

Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. *Efficient BackProp*, pages 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_3. URL `https://doi.org/10.1007/978-3-642-35289-8_3`.

L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22(4):551–575, August 1977. ISSN 1558-2523. doi: 10.1109/TAC.1977.1101561. Conference Name: IEEE Transactions on Automatic Control.

Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. In *International Conference on Learning Representations*, 2020.

V A Marčenko and L A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, apr 1967. doi: 10.1070/SM1967v001n04ABEH001994. URL `https://dx.doi.org/10.1070/SM1967v001n04ABEH001994`.

D. L. McLeish. Functional and random central limit theorems for the Robbins-Munro process. *Journal of Applied Probability*, 13(1):148–154, March 1976. ISSN 0021-9002, 1475-6072. doi: 10.2307/3212676.

OpenAI. `https://github.com/openai/gpt-2/blob/master/src/model.py`, 2024.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR, 2019.

Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra, 2020. URL `https://arxiv.org/abs/2008.11865`.

Mary Phuong and Marcus Hutter. Formal algorithms for transformers. *arXiv preprint arXiv:2207.09238*, 2022.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951. ISSN 0003-4851, 2168-8990. doi:

10.1214/aoms/1177729586. Publisher: Institute of Mathematical Statistics.

Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Statistics & Probability Letters*, 81(5):592–602, 2011. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2011.01.004. URL https://www.sciencedirect.com/science/article/pii/S0167715211000113.

Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond, 2017a. URL https://openreview.net/forum?id=B186cP9gx.

Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks, 2017b. URL https://arxiv.org/abs/1706.04454.

Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pages 3515–3530. PMLR, 2022.

Rylan Schaeffer, Mikail Khona, Zachary Robertson, Akhilan Boopathy, Kateryna Pistunova, Jason W. Rocks, Ila Rani Fiete, and Oluwasanmi Koyejo. Double descent demystified: Identifying, interpreting & ablating the sources of a deep learning puzzle, 2023. URL https://arxiv.org/abs/2303.14151.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. URL http://jmlr.org/papers/v19/18-188.html.

Daniel W. Stroock and S. R. Srinivasa Varadhan. *Multidimensional Diffusion Processes*. Springer, Berlin, Heidelberg, 1997. ISBN 978-3-662-22201-0 978-3-540-28999-9. doi: 10.1007/3-540-28999-2.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018. ISBN 978-1-108-41519-4. doi: 10.1017/9781108231596.